The influence of (technical) developments on the concept of personal data in relation to the GDPR

Bart van der Sloot, Sascha van Schendel & César Augusto Fontanillo López

TILBURG UNIVERSITY

Understanding Society

Table of Contents

Glossary 1
Abbreviations 6

English summary 8

Nederlandse samenvatting 35

Chapter 1: Study design 65
1.1 Introduction 65
1.2 Problem statement 65
1.3 Research questions 67
1.4 Methodology 68
1.5 Themes and topics discussed in this report 70
1.6 Overview report 72
Chapter 2: Anonymization, de-anonymization and non-personal data 74
2.1 Introduction 74
2.2 Legal regulation 74
2.3 Technical developments 86
2.4 Analysis 98
Chapter 3: Aggregation and composition 100
3.1 Introduction 100
3.2 Legal regulation 100
3.3 Technical developments 106
3.4 Analysis 110
Chapter 4: Pseudonymization and de-pseudonymization 112
4.1 Introduction 112
4.2 Legal regulation 112
4.3 Technical developments 116
4.4 Analysis 131
Chapter 5: Sensitive and non-sensitive personal data 133
5.1 Introduction 133
5.2 Legal regulation 133
5.3 Technical developments 140
5.4 Analysis 143
Chapter 6: Analysis 146
6.1 Introduction 146
6.2 Summary of main findings 146
6.3 Regulatory objective of data protection law 150
6.4 Bottlenecks and dangers of under- and over-regulation 154

6.5 Regulatory alternatives 156
6.6 Scenario's 161
6.7 Answers to the research questions 170
Chapter 7: Annexes 175
WODC supervision committee 175
7.1 Interview reports 176
7.2 Workshop report 210
7.3 Provisions in GDPR 213

The influence of (technical) developments on the concept of personal data in relation to the GDPR

Research conducted by
Tilburg Institute for Law, Technology and Society of Tilburg University

Authors
Bart van der Sloot, Sascha van Schendel & César Augusto Fontanillo López

Conducted on behalf of
Wetenschappelijk Onderzoek- en Documentatiecentrum (WODC)

©2022; Tilburg University. Auteursrechten voorbehouden. Niets uit dit rapport mag worden verveelvoudigd en/of openbaar gemaakt door middel van druk, fotokopie, microfilm, digitale verwerking of anderszins, zonder voorafgaande schriftelijke toestemming van de opdrachtnemer.

Glossary

1. Aggregation: anonymization technique aimed at gathering individual-level data and expressing it in summary form.
2. Aggregation based on cryptography: aggregation based on cryptographic primitives, such as secret sharing and fully homomorphic encryption.
3. Aggregation based on data perturbation: aggregation usually performed by noise addition.
4. Aggregation based on third parties: aggregation performed by trusted third parties who collect, aggregate, and transfer the resulting data to authorized recipients.
5. Anonymisation: the process of changing documents into anonymous documents which do not relate to an identified or identifiable natural person, or the process of rendering personal data anonymous in such a manner that the data subject is not or no longer identifiable.
6. Anonymous information: information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.
7. Asymmetric encryption: encryption consisting of the use of two cryptographic keys known as Public Key and Private Key to encrypt (encode) and decrypt (decode) data.
8. Biometric data: personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data.
9. Data: any digital representation of acts, facts or information and any compilation of such acts, facts or information, including in the form of sound, visual or audiovisual recording.
10. Data concerning health: personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status.
11. Data swapping: perturbative masking technique aimed at transforming an original database by exchanging identifiers among individual records.
12. Data synthesis: anonymization technique aimed at generating a dataset which consists of randomly simulated records that do not directly derive from the original dataset while preserving the statistical properties of the original dataset.
13. Dissemination to the public: making information available, at the request of the recipient of the service who provided the information, to a potentially unlimited number of third parties.
14. Document: any content whatever its medium (paper or electronic form or as a sound, visual or audiovisual recording); or any part of such content.
15. Dynamic data: documents in a digital form, subject to frequent or real-time updates, in particular because of their volatility or rapid obsolescence; data generated by sensors are typically considered to be dynamic data.
16. Electronic communications data: electronic communications content and electronic communications metadata.
17. Electronic communications content: the content exchanged by means of electronic communications services, such as text, voice, videos, images, and sound.
18. Electronic communications metadata: data processed in an electronic communications network for the purposes of transmitting, distributing or exchanging electronic communications content; including data used to trace and identify the source and destination of a communication, data on the location of the device generated in the context of providing electronic communications services, and the date, time, duration and the type of communication.

19. Encryption: pseudonymization technique by which data is converted into secret code that hides the information's true meaning.
20. Full data synthesis: data synthesis where every identifier for every record has been synthesized.
21. Fully homomorphic encryption: encryption that permits the performance of extended computations on the encrypted data without decrypting it.
22. Generalization: non-perturbative masking technique aimed at reducing the granularity of the data granularity so that the generated dataset is less precise than the original dataset.
23. Genetic data: personal data relating to the inherited or acquired genetic characteristics of a natural person which give unique information about the physiology or the health of that natural person and which result, in particular, from an analysis of a biological sample from the natural person in question.
24. Hashing with key or keyed hashing: hashing by adding a secret key that alters the output of the function by producing different pseudonyms for the same input according to the choice of the specific key.
25. Hashing with pepper: hashing with salt by adding a secret key to the 'salt' and storing it separately.
26. Hashing with salt: hashing by a so-called 'salt', or auxiliary random-looking data, that alters the output of the function by producing several pseudonyms for the same input.
27. Hashing: pseudonymisation technique aimed at transforming an input of arbitrary length to a result with a fixed length.
28. Homomorphic encryption: encryption that permits the performance of limited computations on the encrypted data without decrypting it.
29. Hybrid data synthesis: data synthesis where the original dataset is mixed with a fully synthetic dataset.
30. Input data: data provided to or directly acquired by an AI system on the basis of which the system produces an output.
31. *k*-anonymity: privacy model aimed at preventing the re-identification of records based on a predefined set of indirect identifiers, so that the ability to link to other information using the quasi-identifier is limited. A dataset is *k*-anonymous if every combination of identity-revealing characteristics occurs in at least *k* different rows of the data set.
32. *l*-diversity: privacy model aimed at preventing the re-identification of records based on a predefined set of indirect identifiers, so that the ability to link to other information using the quasi-identifier is limited. A dataset is *l*-diverse if, for each group of records sharing a combination of key attributes, there are at least *l* well-represented values for each confidential attribute.
33. Local suppression: suppression technique aimed at removing certain individual identifiers in an original dataset with the aim of increasing the set of records that share a combination of key values.
34. Macrodata: dataset comprised of aggregated data
35. Masking: anonymization technique aimed at inducing a relationship between the original record and the generated record, so that the indirect identifier is masked.
36. Metadata: data collected on any activity of a natural or legal person for the purposes of the provision of a data sharing service, including the date, time and geolocation data, duration of activity, connections to other natural or legal persons established by the person who uses the service.
37. Microaggregation: perturbative masking technique aimed at clustering records of an original dataset into small aggregates or groups of *k* elements, where the average of the values of the group over which the record belongs is published in the released dataset.

38. Microdata: dataset comprised of records related to individual data principals
39. Multiparty computation: pseudonymization technique aimed at permitting a set of parties to jointly compute a function of their inputs while avoiding revealing anything but the output of said function.
40. Noise addition: perturbative masking technique aimed at distorting identifiers in an original dataset by adding random noise.
41. Non-personal data: data other than personal data.
42. Non-perturbative masking: anonymization technique aimed at partially suppressing or reducing the detail or coarsening of an original dataset, so that the generated dataset is a reduced version of the original.
43. Partial data synthesis: data synthesis where only identifiers with a high risk of disclosure are synthesized.
44. Personal data: any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.
45. Perturbative masking: anonymization technique aimed at distorting or perturbating microdata so that the statistical properties of an original dataset are preserved in the generated dataset.
46. Privacy model: the framework that specifies the conditions that a generated dataset must satisfy to keep the disclosure risk of data under control.
47. Processing: any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.
48. Profiling: any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.
49. Pseudonymisation: the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.
50. Quantum Computing: computation whose operations can harness the phenomena of quantum mechanics, such as superposition, interference, and entanglement.
51. Record suppression: suppression technique aimed at removing an entire record in an original dataset.
52. Research data: documents in a digital form, other than scientific publications, which are collected or produced in the course of scientific research activities and are used as evidence in the research process, or are commonly accepted in the research community as necessary to validate research findings and results.
53. Re-use: the use by natural or legal persons of data held by public sector bodies, for commercial or non-commercial purposes other than the initial purpose within the public task for which the data were produced, except for the exchange of data between public sector bodies purely in pursuit of their public tasks.
54. Right to data protection: right to the protection of personal data concerning him or her.

55. Right to privacy: right to protection of a person's private and family life, home and communications.
56. Sampling: non-perturbative masking technique aimed at generating a sample of the original dataset.
57. Secret sharing: encryption scheme by which a dealer distributes shares to parties such that only authorized subsets of parties can reconstruct the secret.
58. Sensitive personal data: personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation.
59. Statistical disclosure control: techniques and mechanisms applied to a dataset to eliminate (or reduce) the risk of disclosing information on individual statistical units. These measures usually modify or restrict the amount of data released.
60. Suppression: non-perturbative masking technique aimed at removing the entire or certain identifiers in an original dataset before its release.
61. Symmetric encryption: encryption consisting of the use of one secret key to both encrypt (encode) and decrypt (decode) data.
62. Tabular data: dataset organized in a table with rows and columns
63. *t*-closeness: privacy model aimed at preventing the re-identification of records based on a predefined set of indirect identifiers, so that the ability to link to other information using the quasi-identifier is limited. A dataset is *t*-close if all of its equivalence classes have a distance between the distribution of a sensitive attribute and the distribution of the attribute in the whole table that is no more than a threshold *t*.
64. Testing data: data used for providing an independent evaluation of the trained and validated AI system in order to confirm the expected performance of that system before its placing on the market or putting into service.
65. Tokenization: pseudonymization technique aimed at replacing identifiers with randomly-generated values, known as tokens, without any mathematical relationship and without altering the type or length of the data.
66. Top and bottom coding: generalization technique aimed at setting top-codes or bottom-codes from the identifiers of an original dataset.
67. Training data: data used for training an AI system through fitting its learnable parameters, including the weights of a neural network.
68. Validation data: data used for providing an evaluation of the trained AI system and for tuning its non-learnable parameters and its learning process, among other things, in order to prevent overfitting; whereas the validation dataset can be a separate dataset or part of the training dataset, either as a fixed or variable split.
69. ε-differential privacy: privacy model aimed at controlling the release of information of queries to a released database by mathematically ensuring that a pair of outputs produced by two neighbouring databases (which are the same except for one user's data) are nearly indistinguishable.

Abbreviations

Definition	Abbreviation
Article 29 Working Party	A29WP
Agencia Española de Protección de Datos	AEPD
Advanced Encryption Standard	AES
Artificial Intelligence	AI
American Online	AOL
Statistics Netherlands	CBS
Court of Justice of the European Union	CJEU
Commission nationale de l'informatique et des libertés	CNIL
Council of Europe	CoE
Centrum Wiskunde & Informatica	CWI
Data Encryption Standard	DES
Deoxyribonucleic Acid	DNA
Differential Privacy	DP
Data Protection Authority	DPA
Data Protection Directive	DPD
European Convention of Human Rights	ECHR
European Court of Human Rights	ECtHR
European Data Protection Board	EDPB
European Data Protection Supervisor	EDPS
European Union Agency for Cybersecurity	ENISA
European Union	EU
Fully Homomorphic Encryption	FHE
General Data Protection Regulation	GDPR
Hypertext Transfer Protocol Secure	HTTPS
Information Commissioner's Office	ICO
Intellectual Property	IP
Internet Service Provider	ISP
Law Enforcement Directive	LED
Secure Multiparty Computation	MPC
National Cyber Security Center	NCSC
National Institute of Standards and Technology	NIST
Pseudonymisation Entity	PE
Privacy Enhancing Technologies	PET
Privacy Preserving Techniques	PPT
European Public Sector Information	PSI
Ribonucleic Acid	RNA
Rivest–Shamir–Adleman	RSA
Statistical Disclosure Control	SDC
Secure Hash Algorithm	SHA
Treaty of Functioning of the European Union	TFEU
Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek	TNO
Eindhoven University of Technology	TuE
United Kindgdom	UK
Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein	ULD

United States	USA
Wetenschappelijk Onderzoek- en Documentatiecentrum	WODC

English summary

Introduction

The General Data Protection Regulation (GDPR) is one of, if perhaps the most important framework for the digital domain in Europe and beyond. It sets rules and standards for the processing of data, lays down obligations for persons and organisations processing data (data controllers) and grants rights to individuals whose data are processed (data subjects). Although adopted in 2016, its origins trace back to the 1970s. The decisive element for the application of the data protection framework was and remains whether the data being processed concern information about an individual (natural person).

Although this determination was relatively easy to make in the 1970s, it has become increasingly difficult over time, especially in light of technological developments, the general availability of technology and the push towards open data. These phenomena have meant that it is increasingly possible to derive or infer personal data from datasets that, prima facie, seem to contain no data of this kind. In turn, this has also meant that the legal status of data is increasingly volatile: since data are shared between parties and the operations performed on datasets differ substantially, a single dataset may be considered to contain personal data for one second and no personal data for the next, or as containing personal data in the hands of party A but no personal data in the hands of party B at the same moment.

In response, the legal regime has expanded the notion of personal data over time. In particular, in 1995, the predecessor of the General Data Protection Regulation, the Data Protection Directive, extended the scope of this notion considerably and, therewith, the number of datasets that fell under its reach. Personal data concerns direct as well as indirect information, which refers to data, through which the identity of a person can be inferred, such as descriptions. Personal data not only concerns identifying data – data that can lead to a specific individual at present – but also identifiable data or, in other words, data that do not currently lead to a specific individual, but that may in future. In order to determine whether a dataset contains identifiable data, all means reasonably likely to link the data to an individual should be taken into account. Finally, it is not necessary to know the identity of a person; if data is used to make a decision about a specific individual whose identity is unknown, the data protection regime still applies.

These legislative changes have meant a substantial expansion of the reach of the data protection regime. At the same time, however, the framework still upholds the notion of personal data as the determining factor when deciding whether the rules contained therein apply. In contrast to the restrictive regime laid down for the processing of personal data, the European Union has also adopted another framework for the processing of non-personal data. The Regulation on the free flow of non-personal data essentially holds that no restrictions should be set, either by the public sector or the private sector, with respect to the free flow of non-personal data. Thus, the legal qualification of whether a dataset does or does not contain personal data means that a regulatory regime of almost 180 degrees difference applies, although the proposed Data Governance Act may complicate matters even further.

Now, once again, there have been important technological and societal developments. Big Data, Artificial Intelligence, Quantum Computing and other techniques make it even easier to infer personal data from aggregated, anonymised or encrypted datasets; the general availability of technologies means that it is even more difficult to determine the future status of a dataset; the continued push for Open Data and the re-use of Public Sector Information means that the legal status of data will become even more volatile. In light of these new challenges, questions arise about how the legal regime should respond. Should the concept of personal data be stretched even further? If so, would that not

mean that all data would be considered personal data in practice? Should the current distinction between personal and non-personal data be kept, but a more restrictive regime be developed for non-personal data? What do these developments mean for other data categories in the General Data Protection Regulation, such as pseudonymous data and sensitive personal data?

Against this background, the research question for this study is: *What effect do current and future technical developments have on the data protection framework and the protection afforded to the different types of data with respect to the anonymisation, pseudonymisation, aggregation and identification of data?*

The sub-questions that help answer this research question are:

Identifiability of data

1. What (technical) means are available to link (anonymous) data back to individuals, and to what extent does the availability of other (e.g. open source) data play a role?
2. What (technical) developments are expected in the coming years with regard to the means to (intentionally or unintentionally) link data back to persons?

Anonymisation and pseudonymisation of data

3. What current and foreseeable technical developments can be used for the anonymisation or pseudonymisation of personal data, and what factors are decisive in this respect?
4. What technical developments in the area of anonymisation and pseudonymisation of personal data can be expected in the coming years?

Identifiability in relation to anonymisation and pseudonymisation

5. From a legal and technical perspective, what can be said about the interpretation of the concept of 'means reasonably likely to be used', and what means can be considered reasonably likely to be used and what factors play a role in this?
6. How do the answers to question 5 relate to developments in current and expected technologies toward achieving anonymisation and pseudonymisation?
7. When is it reasonable to say that data can no longer be linked back to an individual and that the dataset they are part of can be considered anonymous?
8. To what extent is the test for indirect identifiability objectifiable?

Consequences of identifiability and anonymisation and pseudonymisation

9. To what extent and in which cases can there be under-regulation when data are no longer linked to individuals through anonymisation and therefore do not fall within the scope of the General Data Protection Regulation?
10. To what extent and in which cases can there be overregulation when increasing amounts of data can be easily linked to individuals through new techniques, undoing measures of anonymisation and pseudonymisation?

Overarching analysis

11. How will current and future technical developments affect the GDPR and legal protection in a broad sense in the coming period?

Several aspects are relevant to answer sub-questions 1 – 8:

- the various legal concepts and the criteria that define and demarcate them;
- the availability of (open access) data and of data processing technologies – in this respect, the European Union's (EU) push for open data and re-use of data is relevant;
- the current and future technological means for anonymising and de-anonymising, aggregating and de-aggregating, pseudonymising and de-pseudonymising data; and
- the impact of the evolving technological capabilities and expanding data landscape on the viability of current legal concepts and demarcations.

Several aspects are relevant to answer sub-questions 9 – 11:

- the regulatory objective of the data protection framework and the light in which the danger of both under-regulation and overregulation should be assessed;
- the regulatory gaps that emerge from the disconnect between the legal and the technological realm; and
- the alternatives to the current legal framework that can be gained from previous European legislation and legislative proposals, literature and interviews.

To answer questions 9 – 11 and to determine whether there is under-regulation and/or over-regulation, it must be determined what the regulatory objective of the GDPR is and should be. Two matters need to be examined in this regard. On the one hand, it is questionable whether data protection law indeed has the sole or main purpose of protecting natural persons. Several authors point out that data protection law was mainly aimed at protecting objective legal principles and general interests, at least initially. On the other hand, the questions under discussion in the legal literature concern to exactly what extent the protection of natural persons is the best basis for future regulation and whether said protection should be extended to groups or society at large.

For this study, three methodological approaches are used.

1. Doctrinal and legal analysis: four types of legal data distinctions are central: anonymous and personal data, aggregated or statistical data and personal data, pseudonymous and non-pseudonymous data and non-sensitive and sensitive personal data. For this purpose, EU and Council of Europe (CoE) laws, their legislative history and legal interpretation are studied.
2. Literature review.
    a. Descriptive literature: technical literature on (de-)identification technologies and privacy/data protection enhancing techniques is assessed.
    b. Normative literature: legal and regulatory literature that describes the challenges of each data category and/or propose new definitions, perspectives or approaches to the various types of data is assessed.
3. Qualitative research methods.
    a. Interviews: interviews were conducted with experts with different backgrounds and areas of expertise: experts on one specific technique, experts with a wide overview of anonymisation/pseudonymisation techniques, experts from organisations that deploy innovative data applications.
    b. Workshop: a workshop was held at the beginning of this study to identify problems and mismatches between the legal and policy domain on the one hand and the technical and practical reality on the other.

The research runs along the following lines.

The legal regime is assessed on three points.

(1) The current legal regime and the existing definitions and explanations in literature or authoritative opinions are assessed to determine what the existing framework is for evaluating data processing.
(2) The history of the legal is was evaluated for three reasons from the point of view of definitions. First, this shows how the data protection framework has been altered over time in response to societal and technological changes. Second, it provides insight into the logic and rationales behind the current definitions and categorisation: why the definitions are as they are and what they aim to achieve. More generally, attention is paid to the discussion about the overarching rationale of the data protection framework, as this is relevant to potential future changes made to the data protection framework. Third, alternative ways of approaching the regulation of data can be found through the various definitions and delineations of the data categories, especially the variations that have been discussed and contemplated in legislative history yet rejected.

(3) The potential future of the data protection framework is assessed. The technological and societal developments discussed in this study have a considerable impact on the interpretation and effects of the current regulatory framework. That is why an overview of the most important lines of thought regarding the potential for altering the current regulatory framework is provided.

The technological realm is assessed on three points.

(1) A brief overview of the technological developments after World War II is provided in order to show how field is constantly in flux. This description shows the background against which the legal framework has altered over time.
(2) Current technologies are assessed, especially in light of the various legal data categories and the boundaries between them. This description shows that it is increasingly possible to de-anonymise datasets and to infer (sensitive) personal data from one or more aggregated datasets.
(3) A description of technological developments that might change the landscape even further in the future is provided. This shows that, if anything, the lines between the various legal data categories will be blurred to an even greater extent.

Attention is also paid to two societal developments, although these are affected by both legal and technological developments.

(1) The study describes how technologies have become general available over time. This means that increasing numbers of governmental organisations, companies and even citizens have highly advanced technological resources at their disposal. As a consequence, if data is shared between various parties or made publicly available, it is increasingly likely that there will be a party that will operate on it in a way that affects the legal status of the dataset.
(2) The study briefly points to the legal and societal push to make data publicly available. This primarily concerns statistical data, public sector information and non-personal data. Mostly, these datasets will not contain personal data in and of themselves, but when combined with other datasets, they may be used to generate (sensitive) personal data. In addition, given the advancement and general availability of technologies, it is increasingly likely that there will be a party that will invest enough resources to de-anonymise or reidentify a dataset.

This study engages with four legal data categories that are engrained in the General Data Protection Regulation: anonymised data, aggregate/statistical data, pseudonymised data and sensitive personal data. Below, a summary of the main findings is provided on the following points:

(1) the current regulation of the various data categories;
(2) the two sometimes conflicting approaches to data regulation that run through the data protection framework;
(3) the general availability of technology and the push toward open data and the re-use of public sector information;
(4) the impact of the changing technological landscape on the regulation of data;
(5) the gaps that exist between the current regulatory regime and evolving technological realities;
(6) the alternatives to the current regulatory regime that are suggested in literature and elsewhere to close these gaps;
(7) the overarching regulatory objective of the data protection framework, in light of which potential changes should be assessed;
(8) the dangers of over- and under-regulation caused by the mismatch between the legal and the technological realm, and
(9) the potential ways forward to solve the existing gaps between the two realms.(10) Once these points are covered, the research question and sub-questions are answered.

1. Legal categories and their elements

This study focuses on four data categories under the data protection regime. In addition to personal data, the study considers anonymous data, aggregate/statistical data, pseudonymous data and sensitive personal data.

Anonymous data

This study assesses the boundary between personal data and anonymous data. Anonymisation means stripping a dataset of direct or indirect identified or identifiable data. If data are properly anonymised, the GDPR does not apply, but the Regulation on the free flow of non-personal data does. From the formal definition of personal data (Article 4 sub 1 GDPR), the relevant recitals (14, 26, 27 and 30) and the subsequent interpretation by the Court of Justice of the European Union (CJEU) and Article 29 Working Party, at least four points stand out.

1. Not only direct identifying data, but also indirect identifying data, and not only identifying data, but also identifiable data will qualify as personal data. The latter means that the current status of data is not determinative; in order to determine the legal categorisation (whether something is personal data or not), account should be made of its likely future status. As the Working Party 29 underlines, although identification may not be possible with all the means likely reasonably to be used today, in a year or ten years, it may. Thus, if data are kept for ten years, the controller should consider the possibility of identification that may also occur in the ninth year of their lifetime. This has significant implications, particularly for open data, which stays online permanently and will be used by various parties.
2. To determine whether data are personal, account should be made of all means reasonably likely to be used for identification. In order to determine the likelihood of identification, an eye should be kept on the costs and the amount of time required for identification, the available technology at the time of processing and future technological developments. Although these are in and of themselves objectively verifiable criteria, their interpretation, as both the Article 29 Working Party and the CJEU have underlined time and again, depends on the context.
3. The question is not only whether the data controller themselves can derive personal data from a dataset currently or in the future, but also whether anyone likely to have access to that data can do so. This, again, is particularly important when data are made available online or shared with multiple parties, as the more parties have access to a database, the greater the likelihood that someone would infer personal data from the dataset, while simultaneously, the means for verifying whether anyone did so, when and why decline.
4. Identification is not required; singling out a person is sufficient. Thus, if an internet company does not know who a person is but can show personalised advertisements to account 87&^%11!, this is enough for the data to qualify as personal data. Likewise, this is the case if an insurer rejects an application from any person from an area with a specific postcode (without knowing their name). On a related note, the Article 29 Working Party has emphasised that data can be considered to 'relate' to an individual because the use of that data is likely to have an impact on a person's rights and interests, taking into account all the circumstances surrounding the exact case. This underlines that it is not necessary that the potential result be major.

Aggregate data

Through aggregation, data can be rendered anonymous by treating the data no longer at the level of $n = 1$, but on the level of $n = 20$, $n = 100$, etc. The analysis of aggregated data may result in information such as – in the most basic terms – of 100,000 people with a green car, 34% have a white couch in their living room. In principle, these data are not considered personal data. However, when parties act on

aggregate data in a way that has a direct impact on natural persons, they may, such as if a car company sends advertisements for white couches to all people that have bought a green car. Throughout the GDPR, there are references to statistical data and aggregate data intended for research purposes. Keeping an eye on public interests (e.g. statistical analysis by National Statistical Agencies is essential for information-based policy-making by governments), the GDPR encourages aggregating data and facilitates statistical research. If data are aggregated to such an extent that no individual data can be extracted nor used in such a way that directly impacts concrete individuals, the GDPR does not apply. In that case, the rules for processing statistical data may apply, which entail standards for confidentiality and safety, inter alia.

When personal data is used for statistical processing, the GDPR applies, but it leaves room for exemptions on the national level. Article 85 GDPR allows for exemptions to the processing of personal data in terms of the freedom of expression. Article 86 GDPR holds that personal data in official documents held by a public authority, public body or private body for the performance of a task carried out in the public interest may be disclosed by the authority or body in accordance with Union or Member State law in order to reconcile public access to official documents with the right to the protection of personal data. Article 89 GDPR provides that the Member States may adopt exemptions – in particular to data subject rights – if personal data are processed for archiving purposes in the public interest, for scientific or historical research purposes or for statistical purposes. There are no factors set out to determine if a database is aggregated to such an extent that it qualifies as non-personal data. This depends on the circumstances of the case, taking into account the general elements discussed previously.

Pseudonymous data

The GDPR is applicable to pseudonymised data, but some exceptions apply to the obligations of data controllers when they process pseudonymised data. Pseudonymisation is regarded as one way to implement technical, organisational and security standards, specific obligations set out by the data protection framework. The GDPR (Article 4 sub 5) defines ‘pseudonymisation’ as the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person. Recital 28 makes clear that pseudonymisation of personal data can reduce the risks to the data subjects concerned, which is why, as set out in recital 29, pseudonymisation is stimulated by the GDPR.

The concept of pseudonymisation is new in the GDPR; it did not have a role in previous data protection regimes. Although the Regulation emphasises that other techniques for ensuring the safe and secure processing of personal data are not excluded by the fact that pseudonymous data are defined separately, it does give this technique a special status. What makes the correct interpretation of this legal concept more complex is that the GDPR often mentions pseudonymisation alongside encryption, which is not separately defined. The Article 29 Working Party supports the increased use of pseudonymisation techniques, of which it distinguishes between five important kinds, including encryption, thus treating certain forms of encryption as a subset of pseudonymisation techniques.

Sensitive personal data

Sensitive personal data are defined separately from ‘ordinary’ personal data under the data protection regime. These are data that are clearly defined and demarcated, and the processing of which is considered to be potentially harmful to the interests of natural persons by definition. In principle, the processing of sensitive data is prohibited, although there are many exceptions that apply to that prohibition. Sensitive data are defined as personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic

data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation. The processing of criminal data by law enforcement authorities is covered by the Law Enforcement Directive.

The CJEU has provided a broad interpretation of what should be considered sensitive personal data. For example, in the *Lindqvist* case, a person had written on a blog that a colleague was on half-time on medical grounds because she had injured her foot. The question of whether having injured one's foot qualifies as 'medical data' was only answered by the Court in a brief, staccato and confirmative manner. Another case, that of *V.*, concerned the transfer of a medical file within the context of employment. The Tribunal pointed out that medical data are particularly sensitive data, thus seemingly making a hierarchy between various categories of sensitive personal data, with medical data at the top.

2. Legal regime: the categorical and the contextual approach

There is a tension between two regulatory approaches in the data protection realm: a contextual and a categorical one; an approach that takes into account the circumstances of the case and an approach that is based on fixed definitions and clear regulatory rules attached to those definitions. Each of these approaches has clear benefits and disadvantages. The first is able to consider all relevant aspects for scenario; it is more adaptive to changing circumstances so does not run the risk of becoming outdated or being circumvented. However, fluid and contextual regulatory approaches have the disadvantage that they are vague and provide little legal certainty, both to the data controller and to the data subject. The second approach solves this problem: it provides a clear set of definitions and categories and attaches to those a clear set of rules. However, the disadvantage is also clear: it runs the risk of being circumvented, becoming outdated and is less granular than a contextual approach.

A deep ambivalence runs through the regulatory approach to data protection on this point.

On first sight, the categorical approach is most apparent. For example, the disconnection of the right to data protection from the right to privacy had to do with a de-contextualisation of the right. In the human rights framework, a claim is assessed on both the *ratione materiae* (does the matter at hand fall under the material scope of the article invoked?) and the *ratione personae* principle (can the applicant claim to be a victim?). With respect to the second principle, there is a significant threshold, as applicants must be able to show that they have suffered from direct, individualisable and substantial harm. Under the data protection framework, both principles are merged. This means that any processing of personal data, however mundane and small, is considered personal data processing to which the GDPR applies. Consequently, the contextual or harm-based element that is essential to evaluations of human rights is omitted from the data protection regime. The application of the data protection regime, which is different from the right to privacy, for example, does not depend on the question of whether there has been harm inflicted on a claimant or rights bearer.

It is also clear whether the data protection framework works with a binary distinction between personal and non-personal data. The EU has provided personal data with the highest form of legal protection in the world through the GDPR and the Law Enforcement Directive (LED), while the EU explicitly discourages restrictions set by private and public sector organisations through the Regulation on the free flow of non-personal data with respect to processing non-personal data. Although the proposed Data Governance Act may complicate this picture, for now, because the distinction between personal and non-personal data is a binary one, the question of whether a dataset is categorised as either one will mean a regulatory difference of 180 degrees. A binary approach can also be seen in both pseudonymous data and sensitive personal data: data are either pseudonymous or they are not, personal data are either sensitive or they are not. With respect to the latter type of data, the categorical approach is even more apparent. The GDPR contains a limited and exhaustive list of types of data that are considered sensitive. The processing of such data is, in principle, prohibited.

A final point that should be emphasised is that the data protection framework as a whole is based on binary distinctions and marked by a categorical approach. For example, it sets out clear differences between various actors, such as the data controller, the data processor and the data subject. Each of these actors has a clearly defined role, set of obligations, rights and regulatory responsibilities. A party cannot be a data processor and a data controller at the same time with respect to the same data processing operation: it is either or. Similarly, with a data processing operation, a party is either a (joint) controller or a data subject.

On the other hand, a contextual approach can be seen. For example, although the distinction between personal and non-personal data is binary, the definition of personal data includes a contextual aspect. The notion of 'identifiable' means that data that does not allow for the identification of an individual at this moment in time but will in the future should be considered personal data now.

Furthermore, although the category of pseudonymous data is in itself binary – data are either pseudonymous or they are not, legally speaking – it is seen by many as an intermediate category between personal and non-personal data. Pseudonymous data are not anonymous, and thus the GDPR applies. However, it is not easy to connect them to an identified individual, which is why the GDPR allows for a number of exceptions when data are pseudonymised. Similarly, although the distinction between non-sensitive and sensitive data is often presented as absolute, all the different rights and obligations apply to both sensitive and non-sensitive personal data alike. The only difference is the legitimate ground for processing the data (Article 6 and Article 9); although Article 9 takes processing sensitive data being prohibited as a starting point, it lists a high number of exceptions to this prohibition, making the difference between the processing of sensitive and non-sensitive data less binary than it seems prima facie.

When the data protection regime applies, most obligations and requirements are context-dependent, meaning that, in general, the more data are gathered, the more sensitive those data are; the higher the risk entailed with data processing or the more parties involved, the stricter the rules and obligations should be interpreted. This contextual approach applies to the obligation to implement a data protection policy, adopt technical and organisational security measures, and technically embed data protection choices by default and by design. At the same time, the obligations and requirements set out in the GDPR have certain context-dependent limitations. For instance, the documentation requirement does not apply to small organisations that do not engage in risky processing operations; a data protection impact assessment only needs to be executed where potential harm is likely; private sector organisations only need to appoint a data protection officer if their core activities consist of regular and systematic monitoring of data subjects on a large scale, or they engage in large scale processing of sensitive data; the data breach notification depends on the harm likely resulting from the breach. Consequently, the core of the data protection framework is highly contextual.

Finally, it has to be emphasised that although the European approach to privacy and data protection is often contrasted with the American one, with the former adopting an omnibus approach and the latter taking a sector approach, the contrast is less sharp than often imagined. The EU explicitly distinguishes between two contexts when it applies data protection rules: the general context, covered by the GDPR, and the law enforcement context, to which the LED applies. In addition, the GDPR promotes the use of codes of conducts, through which sectors can adopt their own interpretations and specifications of the data protection regime. The fact that this possibility has not gained ground, inter alia, because sectors fear the administrative burden of performing oversight and handling complaints does not mean that this is not encouraged by the GDPR.

3. The impact of the availability of data and data technologies on the legal regulation of data

The availability of data is growing exponentially. The data landscape has radically changed between the 1970s, when the first bigger databases emerged, and now, fifty years later. It is not only that more data are being gathered and made available, but more fundamentally, society has changed from an analogous to a datafied one, in which almost all activities are tracked with sensors, cookies, cameras and satellites. Not only do governments use surveillance techniques to monitor the various domains of life, but so do large internet companies and, increasingly, many businesses are being built on datafied and personalised business models. Citizens also have access to all kinds of spyware, drones and other sensory products to collect data about themselves and others. These data are shared via intermediary platforms, stored in the cloud, and made available on closed or open platforms. Another trend is that, with Web 2.0, user-generated content (social networks) has exploded, and hence users themselves have become a major source of personal data, concerning both themselves and their friends.

There is also a legal push to disclose data. In the European Union, there are several laws that require parties to open up. The Open Access Directive, for example, suggests that the Member States make as much public sector information publicly available, free of charge, in open access and in a reusable format as they can. The Company Law Directive requires Member States to take necessary measures to ensure compulsory disclosure by companies of, inter alia, the instrument of the constitution, the statutes, the appointment and the termination of office. The Regulation on the free flow of non-personal data, to provide a final example, dissuades both public sector and private sector organisations from privatising non-personal data.

Three important developments have taken place in recent years with respect to open data:

- Digitisation: government documents used to be available in archives, libraries or specially designated information centres. Nowadays, more and more documents are being made available online. This has an important effect on what is called 'practical obscurity'. The fact that, in the past, one had to make the effort to go to where the documents were stored, request them and view them meant that, in practice, only a limited number of people would access the information. Broadly speaking, these were journalists, historians, critical citizens closely following the government and lay historians researching their family trees. By making the documents public on the Internet and not setting access barriers, anyone can view these documents with ease.
- Active disclosure: in the pre-digital age, most documents were 'passively disclosed'; citizens, journalists and others were given access to specific documents upon request. They thus already had to have a rough idea of what they were looking for, the disclosure of documents required their initiative and the documents were usually only made available for a certain period of time. Currently, documents are increasingly being disclosed actively; governments publish documents not upon request, but at their own initiative. This means that there is no longer a specific reason for which a document is made available. Anyone may access it and at any time.
- Technologies: the technical possibilities of searching through such documents have increased considerably. These include algorithms and AI that can analyse texts for words, correlations, and topics. Whereas previously, it was primarily individuals that sought access to government documents, currently, they are tech companies that are best placed to scan and analyse the millions of governmental documents that appear online.

In addition, given the general availability of data and data technologies, the ease of data gathering and processing and the reduced costs an important shift in the type of data processing operations have taken place. Given the costs and practical and technological limitations for gathering data, many data operations, even up until 20 years ago, used to be targeted. There was a specific and pre-set goal for which specific data on specific entities were gathered. However, currently, many if not most data processing operations concern structural and systemic collections of data, such as cameras and sensors that monitor everyone everywhere in the public domain permanently, mass surveillance operations, and ubiquitous online tracking. This shift means that the data gathered often do not concern pre-identified individuals but groups, categories or the population at large. This, in turn, has initiated a

shift from the analysis of individual data towards that of statistical and aggregated data, from direct to inferred data and from determined to probabilistic data processing.

These developments have an effect on how current data protection legislation is structured and, in particular, on the categorical approach. Three points have been identified.

1. It is argued that working with well-defined and delimited definitions of different types of data only works if a datum falls into one category in a relatively stable way. This is increasingly less so. The nature of the data in 'Big Data' processes is not stable, but volatile. A dataset containing ordinary personal data can be linked to and enriched with another dataset to derive sensitive data. The data can then be aggregated or stripped of identifiers. Subsequently, the data can be de-anonymised or integrated into another dataset in order to create personal data. All this can happen in a split second. The question is, therefore, whether it makes sense to work with well-defined categories if the same 'datum' or dataset if it literally takes one second for it to fall into a different category.
2. It is also increasingly difficult to determine the status of data precisely. The assessment of whether data allow for the identification of an individual and whether the information can be considered anonymous or not depends on the circumstances of the case. Therefore, in order to determine the current status of a datum or dataset, the expected future status of the data must be taken into account. Given the general availability of technologies and the minimal investment required, it is increasingly likely that when a database is shared or otherwise made available, there will be a party who will enrich it with other data. It is thus increasingly likely that if an anonymised dataset is made public, there will be a party that will de-anonymise it or combine it with other data to create personal profiles; that if a set of personal data is shared, there will be a party that will use those data to create a dataset containing sensitive personal data; and so on. On the other hand, there will be other parties who have access to that data but will not engage in such activities; parties who will not use the data, use it as it is provided or even de-identify a database containing personal data. Who will do what is not clear in advance. The legal category to which the data belongs is therefore no longer a quality of the data itself, but a product of a data controller's efforts and investments.
3. The question is whether the distinction made between different categories of data is still relevant. The underlying rationale is that the processing of personal data has an effect on natural persons, while the processing of non-personal data does not, and that the processing of sensitive personal data may have very significant consequences – greater than the processing of 'ordinary' personal data normally has – so that the latter is subject to the most stringent regime, personal data fall under the 'normal' protection regime, and the processing of non-personal data is not subject to any restriction. The question is to what extent this rationale is still tenable in the 21st century. Modern data processing based on aggregate data can have significant individual and social consequences. Similarly, profiling targets groups rather than individuals, meaning that the consequences may be significant, but they may not always be directly relatable to individuals.

4. The impact of current and future data technologies on the legal categories

This study focussed on the technological developments with respect to four fields of application, namely anonymisation, aggregation, pseudonymisation and the inference of sensitive data from non-sensitive (personal) data. The findings with respect to each of these are summarised below.

Anonymisation

The most relevant anonymisation techniques for the purposes of this study are:

1. Masking: aims to generate a relationship between the original record *X* and the generated record *Y,* so that the indirect identifier is masked, which creates anonymity as a result.

1.1 Non-perturbative masking: partial suppression or reduction of detail or coarsening of the original dataset *X*. As a result, dataset *Y* is not a perturbed dataset per se, but rather a reduced version of dataset *X*. Non-perturbative masking encompasses inter alia:
    1.1.1 Sampling: release of a sample *S* of the original dataset *X*. Sampling is suitable for qualitative identifiers which arithmetic operations cannot be done upon, such as the eye colour of an individual or the months of the year;
    1.1.2 Generalisation: reduction of data granularity so that dataset *Y* is less precise than dataset *X*. This technique is appropriate for qualitative identifiers, as it supports the disguise of records with unusual combinations;
    1.1.3 Top and bottom coding: a special case of generalisation by which top-codes or bottom-codes are set from the original identifiers of dataset *X*;
    1.1.4 Suppression: removal of the entire or certain identifiers in dataset *Y* before its release. Since the recovery of information is not possible, suppression is considered the strongest anonymisation technique.
1.2 Perturbative masking: the distortion or perturbation of microdata so that the statistical properties of the original dataset *X* are preserved in dataset *Y*. Perturbative masking encompasses inter alia:
    1.2.1 Noise addition: masking of identifiers by adding random noise;
    1.2.2 Data swapping: exchanging identifiers among individual records;
    1.2.3 Micro-aggregation: clustering of records of dataset *X* into small aggregates or groups of *k* elements, where the average of the values of the group over which the record belongs is published in dataset *Y*.
2. Synthetic data: aims to create a dataset *Y*, which consists of randomly simulated records that do not directly derive from dataset *X* while preserving the statistical properties of the original dataset *X*. As such, standard deviations, medians, linear regression and other statistical techniques can be used to generate synthetic data.

Ways to define anonymity from a technical perspective include, but are not limited to:

1. *k*-anonymity: seeks to prevent the re-identification of records based on a predefined set of indirect identifiers. A cell in a database refers at least to *k* individuals;
2. *l*-diversity: aims to ensure that each group of sensitive identifiers contains different values and that none of these values dominates in terms of frequency of appearance;
3. *t*-closeness: proposes the use of a relative tool to measure the variability of the values of the sensitive identifiers, thus limiting the information gain about the data subjects. All values assumed by the sensitive attribute are considered as equally sensitive;
4. ε-differential privacy: the data controller generates anonymised views of a dataset while retaining a copy of the original data. Thus, those views or subsets are anonymous, but the data controller often still holds identifying information.

Although each of these techniques can be valuable in terms of anonymisation, none of them can guarantee absolute anonymity. Given enough time, resources and adequate technology, practically all anonymised data can be de-anonymised. Even in 2009, Paul Ohm concluded that data can be either useful or perfectly anonymous but never both. Technical literature underlines that this point is true now more than ever. When asked in interviews, technical experts do not expect revolutionary new developments in terms of anonymisation or de-anonymisation, but generally believe that full anonymisation, certainly as understood under the data protection regime, will become increasingly difficult given the general availability of technologies and the general availability of data.

Aggregation

Through aggregation, data in a dataset are presented not at an individual level ($n = 1$), data are presented at an aggregated level ($n = 10$; $n = 100$; $n = 1{,}000$). The higher the level of aggregation, the

more likely it is that the dataset will be considered to contain no personal data from a legal perspective, although such an assessment always depends on the circumstances of the case. The most relevant aggregation techniques for the purposes of this study are:

- Aggregation based on third parties: trusted third parties may collect raw data, aggregate these data and transfer the resulting data to authorised recipients. In this way, recipients only have aggregated data. However, this might not be the case for a trusted third party.
- Aggregation based on data perturbation: random noise is added to the collected data so the original data is not traceable, but aggregated values may still be calculated with a small or negligible error. The drawback of data perturbation is the difference between the original data and the perturbed data, which may lead to disparities in the computation in certain cases.
- Aggregation based on cryptography: cryptographic primitives can be used to overcome the drawbacks of the previous methods. Fully homomorphic encryption is an encryption technology that allows the performance of analysis in the ciphertext in the same way as in the plaintext without sharing the secret key. This implies that the computation is performed over the encrypted data without the need to decrypt it, thus enabling data sharing with third parties. The results of the computation are equally encrypted, so only the exporters of data are able to decrypt it.
- Statistical Disclosure Control: Perhaps the most important technique in terms of aggregating data, especially in light of disclosing the data, is Statistical Disclosure Control (SDC). SDC aims to eliminate identifying information in a dataset, both directly and indirectly, while preserving data quality as much as possible. The specialist in charge of protecting the data has to use different disclosure control methods in such a way that the minimum required level of protection is achieved, and that the information loss is as small as possible, which will differ per situation. What constitutes information loss cannot be determined as such, as information is a subjective term that can be defined differently by each user.

Although the technical opportunities for anonymising data in aggregated datasets are high, and in general higher than when data are not aggregated, a new problem emerges, which is identified in technical literature as the composition problem. This means that personal data may be inferred from the combination of two or more datasets not containing any personal information themselves. This may concern data about identified people that used to be in those databases, but the data can also regard other people. In addition, it should be emphasised that if a party used general information to make decisions that affect individuals, this would qualify as personal data in the legal realm. Obviously, however, it is difficult to assess which party will use which aggregated data for what type of decision-making beforehand.

Although anonymisation of aggregate data is potentially possible in isolation (e.g. only taking the dataset as a relevant resource for identification purposes), both literature and experts interviewed for this study agree that this will be always less determinative. This is not so much due to evolving techniques, but it has to do with the expanding data landscape and the availability of open data. Because it is likely that almost any aggregated dataset will be used for inferences on a personal level, in time, for compositional activities and/or for developing decision-making policies that have an effect on people, from a legal perspective, no aggregated dataset should qualify as definitely falling outside the data protection regime.

Pseudonymisation

The most relevant pseudonymisation techniques for the purposes of this study are:

1. Hashing is a technique that can be used to derive pseudonyms. In a nutshell, hash functions are functions that compress an input of arbitrary length to a result with a fixed length. This fixed-size output is called a ‘message digest’, ‘hash value’, ‘hash code’ or simply ‘hash’. In this way,

if an identifier *m* is used as an input in the hash function *h,* the function will return a fixed-size pseudonym *h(m)*.

2. Hashing with key or keyed hashing builds on conventional hashing by adding a secret key that alters the output of the function *h*. Hashing with key can produce different pseudonyms for the same input according to the choice of the specific key.
3. Hashing with salt is a variant of keyed hashing, where a conventional hash function together with a so-called 'salt' – or auxiliary random-looking data – is used. Just like keyed hashing, hashing with salt produces several pseudonyms for the same initial identifier. As such, hashing with salt enjoys the same properties as keyed hashing as long as the salt is appropriately secured, and third parties do not have knowledge of it.
4. Peppered hashing consists of adding a secret to the salt during the hashing and storing it separately from the salts and pseudonyms in another medium, for instance, in a hardware security module. The pepper, therefore, shares certain properties with salt in that it is a random value and is similar to an encryption key in that it must be kept secret.
5. Tokenisation consists of replacing identifiers with randomly-generated values, known as tokens, without any mathematical relationship and without altering the type or length of the data. This is an important difference from encryption. As opposed to the latter, the invariability of data types and lengths in tokenisation prevents any unintelligibility of information through processing in intermediate systems. At the same time, it also implies a decrease in the computational resources needed to process the tokens. Since there is no involvement of keys or algorithms to derive the original identifier from the token, the knowledge of a token does not imply the disclosure of personal data.

The most relevant encryption techniques for the purposes of this study are:

1. Symmetric encryption, which consists of the use of one secret key to both encrypt and decrypt electronic information. Parties relying on symmetric encryption must share the secret key to enable the decryption process. Symmetric encryption transforms the initial identifier – and the complete dataset – into a pseudonym (or ciphertext), which is then decrypted to reveal the initial identifier.
2. Asymmetric encryption, which consists of the use of two keys – a public and a private key – to both encrypt and decrypt electronic information. Parties relying on asymmetric encryption must rely on the public key to encrypt the data and on the private key to decrypt it. Public and private keys are mathematically related but appropriately distinguished by the introduction of randomness in the encryption process to prevent the determination of the private key.
3. Homomorphic encryption, which allows computation on encrypted data. Computing on encrypted data refers to the fact that a party P_n, having the initial identifiers or input m_n, and wanting to calculate the function *f* to obtain $f(m_1,\ldots,m_n)$, can instead compute the encryptions or pseudonyms of the inputs c_n to obtain $f'(c_1,\ldots,c_n)$, which can be decrypted to $f(m_1,\ldots,m_n)$. The benefit of homomorphic encryption is that personal data remains confidential while being analysed or mined without the need to decrypt it and compromise the output.
4. Multiparty computation (MPC), which is different from the three previously discussed techniques, although it is related to homomorphic encryption. MPC is a technique that deals with protocols that allow a set of parties to jointly compute a function of their inputs or identifiers while avoiding revealing anything except the output of said function. MPC allows the parties' input to remain secret during the whole processing of data aggregation, so it is considered a sophisticated privacy-preserving tool for pseudonymisation. It can be used as an encryption technique, but it is much broader in terms of its potential application.

Mostly, technical literature describes encryption and pseudonymisation techniques as forms of privacy-enhancing or preserving technologies. Which technique is most appropriate depends on the context, the type of data, the actors involved and other safeguards in place. That is why no one technique can be said to be the preferred option and no technique can be ruled out categorically even though some

techniques are generally deemed weaker than others. Some pseudonymisation or encryption techniques, especially when adopted in combination with other privacy-enhancing technologies, can be so strong that they may provide better protection to data subjects' interests than certain anonymisation techniques.

Inference of sensitive data

It is clear from technical literature that it is becoming ever easier to infer personal from aggregate data as well as sensitive personal data from personal and non-personal data. For example, statistical agencies and census bureaus commonly publish aggregated datasets, which they believe do not contain any personal information. Yet through so-called 'database reconstruction attacks', it is often possible to reconstruct the sex, age, race, ethnicity and fine-grained geographic location reported for about half of the population contained in the dataset. In addition, by combining two datasets that themselves do not contain any personal data, personal data may be gained and even sensitive personal data may be inferred. Consequently, due to both the availability of open data and increased technological capacity to infer data, sensitive data can be distilled from both 'ordinary' personal data and non-personal data. Both the literature consulted and experts interviewed for this study agree that this trend will only increase over time.

5. Gaps between the legal regime and the technological reality

There are several tensions between the evolving technological realm and how the legal framework has been drawn up. The most important ones for the purposes of this study are detailed below.

Anonymous data

1. While the legal framework distinguishes between anonymous and pseudonymous data, for technical experts, this distinction is not uncontested. From a technical perspective, data could be called anonymous if a number of relevant variables are removed. Many technical experts assume levels of anonymity. There is a scale from full anonymity to direct identifiability rather than a binary distinction, as is prevalent in the GDPR.
2. The fact that the GDPR sets no time limit on when data can be re-identified or de-anonymised means that it is highly likely that, at some point in time, the data will be linked to a natural person.
3. A number of technical experts question whether the legal definition of anonymous data can be upheld in the 21st century, as it will be increasingly difficult to meet the legal threshold. From a technological perspective, it is almost impossible to have truly anonymous data. In particular, when anonymised datasets are shared or made available online, it is likely that there will be a party that re-identifies the data or merges it with other datasets to arrive at personal data.
4. Some authors conclude that the state of the art linked to the techniques listed by the Working Party 29 confirms that anonymization methods face big challenges with real data and that it cannot longer be considered from a static perspective, but it requires a dynamic one.
5. A general sentiment that was shared is that the term 'anonymisation' was unclear and vague due to its many open-ended factors. A special point of reference was the term 'reasonably likely'.

Aggregate data

1. Legal regulation treats (micro) data and aggregate (macro) data the same, although there are clearly different risks attached to disclosing micro and macro data. On a record level, to speak of absolutely anonymous data generally requires stripping the dataset to such an extent that virtually no relevant information remains, whereas on the aggregate level, there are many more opportunities to protect individuals from identification, yet aggregated data is likely to be linked to individuals through other ways, especially when made available online.

2. The increased availability of open data makes it hard to do a proper assessment of the risks involved when statistical agencies or other parties release aggregated data.
3. Statistics are used to generate knowledge by analysing existing data to make assumptions about individuals, for example, by mapping past experiences and establishing correlations between certain characteristics and particular outcomes or behaviour. AI and Big Data analytics allow people to be profiled in actionable ways without being personally or individually identified. As the state of technology allows more information to be extracted from non-personal data, a greater role is awarded to the use of such data, whether individuals in it can be identified or not. This trend cannot be adequately addressed under the data protection regime, which is highly invested in the notion of the indefinability of individual natural persons.
4. For many technical experts and professionals, the legal regime gives conflicting signals. On the one hand, open data, the re-use of public sector information and data portability are promoted, and on the other hand, privacy, secrecy and data protection are emphasised. What makes this tension more complex is that legislators and courts do not present a uniform view as to what extent the collection and use of aggregated data should be regulated or to what extent aggregated data can also be personal data.

Pseudonymous data

1. Technical experts traditionally presume that pseudonymisation means replacing one or more identifiers with a pseudonym. However, the GDPR defines pseudonymisation as processing where additional information that allows for re-identification is stored in a different place; there can be pseudonymous data without having an explicit pseudonym. The two definitions are very close, but not fully identical.
2. From a technical perspective, it is not clear why pseudonymisation, as a form of harm prevention, should have a special status within the legal framework, while there are multiple ways and techniques for doing so. Favouring this one technique seems to be in contrast with the supposed technical neutrality of the data protection framework.
3. Not all forms of pseudonymisation are equally safe. The legal regime does not give guidance on which type of technique is most appropriate for what type of context.

Sensitive personal data

1. Experts question the fixed categories of sensitive data used in the GDPR. For example, financial position, socio-economic background and income could be treated as sensitive data in many cases, because the potentially detrimental effects of processing such data might be at least as severe as the processing of the membership of a political organisation or a union. Experts indicate what is or should be considered sensitive personal data varies per region or country, which is why working with one fixed list of types of sensitive data for all EU countries alike may be particularly challenging.
2. The legal regime focuses on fixed categories of data, while what is or is not sensitive from a technological perspective does not depend on the type of data. Data processing can be sensitive and harmful even without the categories of data listed in the GDPR being involved or can be non-harmful even if one or more of the types of data categorised as sensitive are processed.
3. Many technical experts highlight the fact that sensitive information can often be derived from non-sensitive personal information and even from non-personal data. The binary distinction between sensitive and non-sensitive data used in the legal regime does not take sufficient account of the technological complexity and reality on this point.

6. Regulatory alternatives found in law and literature

Alternatives to the current regulatory regime have been suggested. Those most relevant for the purposes

of this study are:

Anonymous data

1. Doing away with the distinction between anonymous and non-anonymous data. If it is increasingly likely that data will be de-anonymised and if non-personal data can be used for high-impact data processes, the choice to place anonymous or non-personal data outside of the scope of data protection law could be redundant.
2. Using a narrower concept than the current definition of personal data to distinguish more clearly between personal and anonymous information. For example, the notion of 'identifiable' could be done away with, or a specific horizon or time limit could be added to it.
3. Creating different levels of identifiability, meaning that the application of the data protection framework is not black and white but gradual. For example, the more data can be said to be anonymised, the fewer data protection standards apply.
4. Instead of working with the highly contextual 'all the means likely reasonably to be used either by the controller or by any other person', working with a phrase that was suggested in the legislative process of the Data Protection Directive could be considered, namely 'at the price of an excessive effort'.

Aggregate data

1. Instead of seeking a compromise between open data and data protection, an option could be to have the data protection regime prevail over or provide the main framework for using, sharing and making public statistical and aggregated data. Essentially, this is what the CJEU proposed in Latvijas Republikas Saeima.
2. An even farther-reaching option could be to return to one of the earlier rules on statistical data, namely that 'statistical data should be released only in aggregate form and in such a way that it is impossible to link the information to a particular person.'
3. An extensive framework could be set up to reconcile the need for open data and processing statistical data on the one hand and the need for privacy and data protection on the other. Such a framework should be adopted at the EU level and not left to the Member States, and it would need to specify how these two principles could be reconciled in concrete situations in detail.
4. The data protection framework could be more explicit in terms of a threshold or boundary for data anonymity when data are aggregated or in terms of the technical standards to be applied when disclosing aggregated datasets.
5. A more radical alternative could be to find ways to base privacy and data protection regulation on concepts other than identifiability. For example, some authors have suggested abandoning the (exclusive) focus on individual privacy and linkability, and instead or in addition focusing on groups, categories and data collectives.
6. More concrete rules for disclosing aggregated data could be developed, such as having a minimum number of people in a cell with a frequency count table or rules on dominance with quantitative magnitude tables. Checks for group disclosure could be stipulated too.

Pseudonymous data

1. The GDPR or the European Data Protection Board could provide more guidance as to which types of pseudonymisation techniques are deemed most suitable for what contexts.
2. Aligning pseudonymisation with the concept of data custodianship could be considered. A data custodian could function as a Pseudonymisation Entity, responsible for processing identifiers into pseudonyms using the pseudonymisation function (which can be a controller or processor), which can allow data access under specific conditions to researchers or companies in an interconnected data ecosystem and shield data against unwanted or unlawful access.

3. Some experts suggest deleting the specific reference to pseudonymisation from the GDPR, both because it is considered too vague and because there is no reason to favour this technique over other risk aversion techniques.
4. Others, on the contrary, have suggested giving the category of pseudonymous data an even more prominent role, making it an official intermediate category between anonymous data and personal data.

Sensitive personal data

1. Several authors have suggested that the sensitivity of data processing no longer depends on the type of data being processed, but rather on the processing technologies and the use they are put to. Consequently, they have suggested omitting the special regime for sensitive personal data from the data protection framework.
2. Broaden the list of sensitive data and include in it, inter alia, financial data, as was suggested when drafting the GDPR but was ultimately rejected.
3. Alternatively, it has been suggested to work with a list of examples rather than with fixed categories, which was the original approach to the regulation of sensitive personal data. Also, the introduction of a residual category could be considered, similar to the reference to 'or other status' in Article 14 of the European Convention on Human Rights (ECHR).
4. The differentiation between the various categories of sensitive personal data could be considered; this is an approach that seems to be taken by the CJEU, placing health data in the most sensitive category, while other data could be considered as less sensitive.
5. While most concerns are about whether the GDPR is strict enough on special categories of data, there are also arguments to consider from the opposite perspective. There is an ongoing discussion concerning the extent to which it is possible to process sensitive personal data in order to prevent discrimination, for example, in AI systems. Using sensitive personal data may be necessary for avoiding discrimination, especially when it comes to data-driven decision making. Thus, in order to further one of the underlying rationales of the category of special data – namely to prevent discriminatory practices – it may be necessary to process more sensitive personal data, instead of less.

7. Regulatory objective of the data protection regime

In order to assess whether there are regulatory gaps and, if so, where they are, it is necessary to assess what the regulatory objective of the privacy and data protection regime actually is. This is a matter of debate: should data protection be seen as essentially putting limitations and restrictions on data controllers or as giving control rights to data subjects?

On the one hand, reference is made to Article 5 of the General Data Protection Regulation, which is seen as the backbone of the law. It holds that personal data should be processed lawfully, fairly and in a transparent manner, collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with these purposes. Processing should be adequate, relevant and limited to what is necessary in relation to the purposes for which the data are being processed, and they should be accurate and, where necessary, kept up to date, kept in a form that permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are being processed and they should be processed in a manner that ensures appropriate security. These are all obligations posed on the data controller, which are applicable independent of any rights being invoked by data subjects. On the other hand, there are increasing numbers of rights attributed to data subjects in the data protection regime, the GDPR. In addition, particularly due to German influence, the notion of informational self-determination has become increasingly popular. Thus, some argue that rather than obligations being imposed on data controllers, the rights of data subjects are the core of the data protection regime.

What complicates the assessment of this matter is the fact that the data protection regime does not only have a protective objective, but Article 1 GDPR also acknowledges the processing of personal data in the EU as objective facilitating. One of the explicit goals of the 1995 data protection framework was to remove obstacles to the transfer of personal data within the Union by laying down one common level of data protection. Before the Directive, each country had different data protection standards, hampering the use and transfer of personal data. Adopting a single EU-wide data protection framework solved this. The rules in the GDPR seldom prohibit specific data processing operations. In most cases, they lay down procedural safeguards and principles that ensure proper data processing operations. Data protection may thus be said to further data processing by laying down a general framework.

Finally, the GDPR contains many explicit exceptions for specific processing operations. Most important for this study are those that relate to freedom of speech, archiving, statistical research, open government and the re-use of public sector information. The EU has made the choice to go beyond promoting openness and transparency vis-à-vis governmental practices; it has stimulated the re-use of government information. Still, the Open Data Directive makes clear that it does not affect the GDPR; what this means in practice is left open.

There is ambiguity within the EU regarding how to deal with conflicts between the various regimes and the rationales underlying them. In general, the EU regulator is set on issuing a regulation that is based on clear and separate data categories while the courts have adopted more contextual and fluid approaches. Advisory bodies, such as the Article 29 Working Party and the EDPB, also propagate a flexible approach and have stretched and broadened the scope of, inter alia, personal data over time. Courts had set clear limits when regulators use data distinctions to adopt lower levels of protection, such as when the CJEU declared null and void the Data Retention Directive. A similar approach can be witnessed with respect to the move toward open data and the re-use of public sector information. While this is highly encouraged by the EU regulator, the courts are more hesitant. For example, the CJEU questioned whether, in order to protect or improve road safety, it was necessary to grant third parties access to data about traffic violations. It found that the regime allowed third parties to access the information even if they had purposes other than those related to increasing road safety, which was not allowed.

8. Dangers of over and under-regulation

The difficulty of assessing the existence of regulatory gaps and the desirability of regulatory alternatives is that the discussion on the regulatory goal(s) of the privacy and data protection framework should be settled first, but it remains a matter of debate – a debate that this study can obviously not settle. In addition, there is no preferred regulatory approach: a categorical, a contextual or a hybrid one. Each has its own advantages and disadvantages. Consequently, it is both a matter of perspective whether regulatory gaps exist and, if so, what this means for the legal protection of data in the broader sense in the future. In addition, choosing between different regulatory options entails a choice of where to put the regulatory prerogative. The more clarity provided in the legal regime, the more the prerogative is put with the legislative branch. The more a contextual approach is taken, the more the correct interpretation of the rules per context has to be given by the judicial and/or the executive branch. The former has the advantage of democratic legitimacy, the latter of practical validity. The former has the advantage of providing legal certainty by applying one approach to all situations alike; the latter has the advantage of being able to provide regulatory granularity.

To give an example, perhaps the essential question this study raises is whether the notion of ‘personal data’ and the sub-criteria of ‘identifiability’ and ‘all means reasonably likely’ should be retained, or whether non-personal data should be provided as a form of protection, for example under a ‘GDPR-light’ regime, or whether a gradual scale to identification should be introduced. That question is

dependent on what the regulatory rationale of the data protection framework is believed to be. If it is providing protection to natural persons' individual interests, then there is no direct need to also regulate the processing of aggregated or anonymous data. In order to tackle potential harms that arise from data policies and practices based on group profiles, it might be left to the courts to interpret the regulatory regime as to cover those harms either on the basis of the GDPR or under Article 8 ECHR. If the regulatory objective is to curtail data power by public and private sector organisations, then it both makes sense to also set limits on and requirements for the processing of non-personal data, and it would be no problem to expand the data protection regime to also cover the processing of non-personal data and disconnect its material scope from the identifiability of a natural person.

Both choices also beg the question of regulatory specificity. The regulator has, so far, maintained a strict regulatory distinction between non-personal and personal data, however, in practice, this distinction is difficult to draw. Courts have consequently expanded the definition of personal data to cover data that is increasingly peripheral to the natural person, while data controllers are asking for more cues on how to make that distinction. The danger of leaving the current approach intact is that responsible data organisations will err on the safe side, while others will stretch the boundaries of the law. In addition, the less regulatory clarity is given, the more difficult it will be to enforce the rules, because every data processing operation may require its own assessment of legality and legitimacy. Consequently, if the choice is made to keep the current regulatory regime intact, the question remains whether more regulatory guidance should be provided to data controllers on how to draw distinctions between data categories.

In addition, if the choice is made to cover non-personal data, again, two different approaches can be taken: a categorical and a contextual one. Either the regulatory regime maintains a separation between non-personal and personal data but attaches a different regulatory regime to non-personal data, or it does away with this differentiation and potentially other data distinctions and makes the type of rules and the regulatory burden put on data controllers fully dependent on the case by case assessment of the risks involved (either related to individual, group and/or societal interests).

Then, regarding overregulation, it matters to what extent stimulating data processing operations is set on the same foot as the protective rationale of the data protection regime and how the goal to promote open data environments and the re-use of public sector information is evaluated. Should the latter rationale be seen as an equally important rationale as the protective rationale, or can this rationale only be furthered within the boundaries set by the protective rationale? If the latter is the case, overregulation is not an important risk, while avoiding under-regulation should be the main objective. If, however, the rationales are considered to be of equal importance, furthering one almost by definition has an impact on the other. Then, the question would be what type of regulation would be most effective. Although a contextual framework seems to leave the most room for data innovation, at first sight, data controllers often call explicitly for more regulatory clarity and certainty, because they fear backlash and investments that will not pay off.

A similar point should be noted with respect to the protective rationale. Experts have stressed that the approach taken by the GDPR, under which processing sensitive data is in principle prohibited, is increasingly missing the goal it sets out to achieve, namely, to protect individuals from harm. In order to prevent discriminatory practices in AI-systems, it may be necessary to process sensitive personal data. Others have stressed that even more generally, it may be necessary to focus not on data minimisation but on data *minimummisation*, i.e. requiring a minimum level of data to be gathered, analysed and stored rather than a minimal level of data be gathered in order to protect individuals vis-à-vis AI systems and their outcomes. Thus, it is a matter of debate whether the protective rationale is best served by laying down limitations on data processes.

9. How will the current and future technical developments affect the GDPR and legal protection in a broad sense in the coming period?

It is clear that current and future technological developments and general data availability are resulting in anonymisation becoming increasingly difficult. The status of data will become increasingly volatile, less owing to the characteristics of the data and datasets themselves, and more due to data controllers' efforts. Legal categories will become more fluid and porous, and one database may be legally qualified differently per party that has access to it. A database that, in isolation, only contains non-personal data one moment may be turned into personal data by being combined with another database the next, which may be used to infer sensitive personal data the next moment, only to be aggregated and anonymised the next moment again. Given these trends and given of the notions of 'identifiability' and 'all the means reasonably likely to be used', more and more data, if not all, will fall under the data protection framework and may even need to be treated as (potential) sensitive personal data, meaning that the strictest of all regimes would apply.

Whether this is deemed problematic is a matter of debate and depends on what the rationale of the data protection framework is deemed to be and what effects of under-regulation and over-regulation are most likely. This study did not find different scenarios for how the technological realm and the availability of open data will develop over time – literature, experts interviewed and experts invited to the workshop held for this study all pointed in the same direction: anonymisation will be increasingly hard, legal categorisation will be increasingly difficult and the quality of the data will be increasingly an effect of the efforts by the data controller. However, several scenarios were found for how the legal regime could respond to the increased availability of open data and the general availability of technology. Five strategies can be deduced from the suggestions, which can be summarised as follows.

1. **Leaving the data protection framework as is:** The data protection framework is regarded as forming a perfect equilibrium between its protective rationale and the rationale promoting data processing operations, between opting for a categorical and a contextual regulatory approach and between leaving the regulatory prerogative to the legislator and allowing judicial and executive authorities to refine concepts and rules in practice, with an eye to specific contexts and situations. Although technological practice may be said to diverge from the regulatory regime and may very well do so more over the years, this does not mean that the rules should change. Rather, more should be invested in ensuring that practice is kept in conformity with the rules. To the extent that processing non-personal data has an important impact, such as is already covered by the GDPR when decisions are taken in which a person is singled out or significantly affected, or by Article 8 ECHR, when policies affect the very broad notion of private life. The European Court of Human Rights has been willing to develop a regime for metadata collection when necessary and has accepted claims in which no personal harm was endured by the claimant, instead focussing on the societal effects of large-scale data processing. Data protection law does not need to solve all problems of the data-driven environment.
2. **Keeping the data protection framework and investing in more precise definitions:** The main outlines and contours of the current regulatory regime are deemed fit for the 21st century, while the main regulatory challenge is the need for further clarity of the definitions of the different data categories, the boundaries between different categories and the regulation of those types of data. Under this scenario, various regulatory alternatives are possible, such as more guidelines being issued and the introduction of a burden of proof on the data controller for showing that data are anonymous and/or encrypted. To provide more clarity on the distinction between non-personal and personal data, the contextual elements in the definition of personal data and in the description of anonymisation could be removed. This would decontextualise the question of whether personal data are processed and whether the data protection framework applies. Also, the category of pseudonymous data could be omitted. This category is critiqued both for its vagueness and because it privileges one privacy-preserving

technique over others, for which no clear explanation exists. Finally, it may be considered to extend the list of sensitive personal data. Potential additional categories that were identified in this study include financial and socio-economic data, data about children, locational data and metadata.

3. **Keeping the data protection framework and investing in more contextuality:** The main regulatory challenge is regarded as the lack of contextuality and adaptability of the current regulatory regime. Again, several regulatory alternatives have emerged during this study, including the addition to the list of Article 5 GDPR of the principle of contextuality, requiring the data controller to consider each principle, obligation and requirement under the data protection framework in light of the context in which the data processing takes place. Alternatively, reformulating the list of sensitive data in the way it was originally formulated could be considered, namely as examples rather than an exhaustive list, or including a residual category similar to Article 14 ECHR. Pseudonymous data could be granted a more prominent position as an intermediate category between non-personal and personal data.
4. **Revising the data protection framework, using clearly defined data categories:** Strategy 4 is similar to strategy 2, but under this scenario, a fundamental overhaul of the current regulatory framework is necessary. Under this scenario, it is believed that it is still possible to work with categories of data, even the current ones, but in light of the technological developments, the regulatory regime applied to them is in need of reconsideration. A number of regulatory alternatives could be considered, such as adopting a GDPR-light regime for non-personal data; this could imply, for example, that all data processing must accord with the principles contained in Article 5 GDPR. Also, potentially in light of a protective regime on non-personal data, structuring the data processing regime around stages of data processing could be considered: gathering and storing data, analysing data and using data or the outcomes of data analysis. The current regulatory regime almost exclusively focuses on the moment at which data are gathered and stored. There are virtually no rules on the analysis of data and no rules on the use of data, perhaps with the exception of one provision on the prohibition on automated decisions making. This is deemed problematic because the core of most current-day processing operations is in the analysing of data. For the analysis of data, inspiration could be sought from the rules applicable to statistical agencies.
5. **Revising the data protection framework, removing clearly defined data categories:** Strategy 5 is similar to strategy 3, but in this scenario, a fundamental overhaul of the current regulatory framework is necessary. In this scenario, it is impossible to work with different data definitions and to attach different levels of regulatory protection to each of those. Instead, a fully contextual approach should be taken, fully dependent on a case-by-case analysis of the potential harm that results from a certain processing operation. This harm could be linked to individual and/or societal interests. Most of the current obligations and requirements could be left intact, yet they would be made dependent on the level of risk and harm. The GDPR could essentially be boiled down to a simple set of rules, namely a list of principles and obligations for data controllers that are currently in the regulations, and specify that these apply to them, taking into account the state of the field, the costs of implementation and the nature, scope, context and purposes of the processing, the nature of the data as well as the risk of varying likelihood and severity for the individual and/or societal interests.

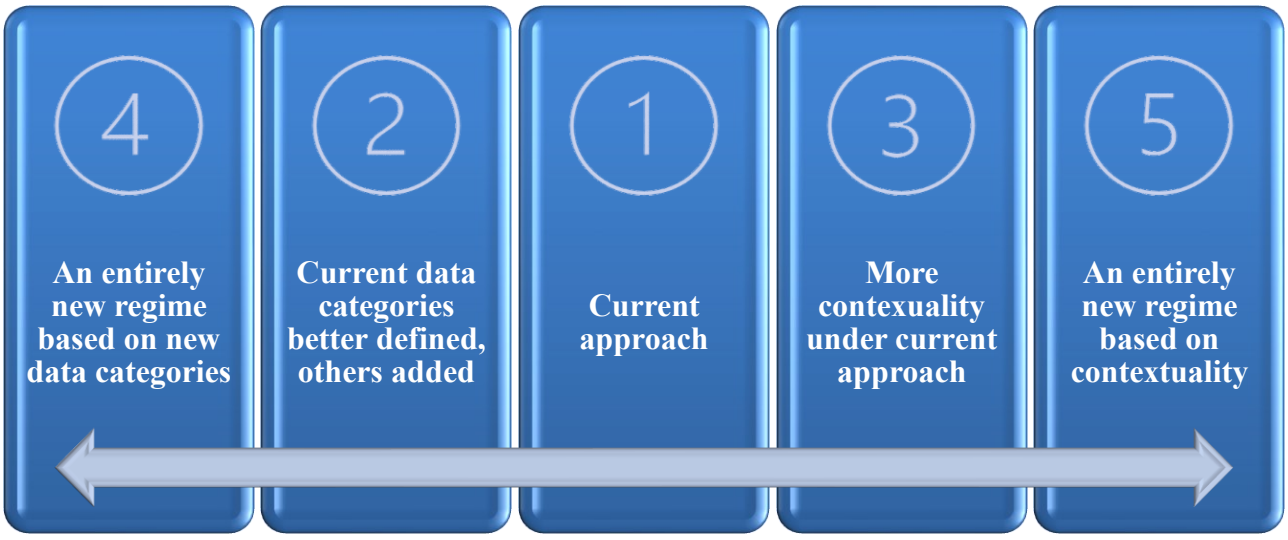

Figure: Scale from a fully categorical approach (option 4) to a fully contextual approach (option 5)

10. Answers to the research questions

1. What means are available to link (anonymous) data back to individuals and to what extent does the availability of other data (e.g. open source) play a role?

There are many means available to link data back to individuals. This study has not arrived at a full and exhaustive list of possibilities, but it does discuss a number of common means for doing so. Examples include: database reconstruction attacks through which an aggregated database is re-identified; composition, through which two or more anonymised datasets merged together can result in (sensitive) personal data and several de-anonymisation technologies. Information may be inferred from anonymised datasets about people that were not in the dataset in the first place and aggregated data, in particular, may be used for decision-making processes, which may have a significant effect on citizens in general and specific groups. If the latter is the case, these data may qualify as personal data.

Open data plays an important role in this respect, so much so that many experts point out that although it may be possible to de-individualise a dataset taken in isolation because it is possible to combine it with other data freely available online, it can never be excluded and, on the contrary, it is increasingly likely that anonymised dataset will be de-anonymised by one party or another in time. Aggregated data, when they are made available, may be used for decision-making that affects specific identified or non-identified citizens. How data will be used cannot be controlled or estimated beforehand with certainty. However, the chance that if data are made available online, they will be used by a party in ways that have an effect on concrete individuals, groups or society at large is increasingly likely.

2. What (technical) developments are expected in the coming years with regard to the means to (intentionally or unintentionally) link data back to persons?

It will be increasingly difficult to ensure (legal) anonymity. Already now, experts interviewed for this study doubt whether it is possible to meet the legal criteria for anonymity. While the legal regime treats anonymity as a binary matter, most technical experts see it as a scale. Most technologies and counter-technologies are involved in a never-ending game of cat and mouse. This is also believed to be the case for the future for, inter alia, anonymisation and de-anonymisation techniques, aggregation and inference techniques and for encryption and decryption. What is the most fundamental shift is the general availability of such technologies. This means, especially when data are made available online, it is increasingly likely that there will be some parties who will use advanced technologies to decrypt, re-identify or de-anonymise data and invest the necessary time, energy and effort for doing so. An

important development with respect to encryption is quantum computing.

Quantum computing possesses certain characteristics derived from quantum mechanics that make it possible to solve complex factorisation problems with which traditional computers struggle. Instead of working with bits, quantum computers work with quantum bits or *qubits*. Qubits are able to simultaneously take a value of 0 or 1, as opposed to traditional bits, which have a single state of 0 or 1. This enables quantum computers to perform multiple parallel calculations for which conventional computers are not suited. As a result, certain authors have claimed an alleged 'supremacy' of quantum computing over conventional computing, which would allow for the cracking of the present cryptography. In addition, if quantum computing becomes operational, it is said to be able to break all or most existing forms of encryption, just like current techniques can decrypt Data Encryption Standard (DES) encrypted messages from 40 years ago.

3. What current and foreseeable technical developments can be used for the anonymisation or pseudonymisation of personal data, and what factors are decisive in this respect?

Various techniques exist for both anonymisation and pseudonymisation. Examples of anonymisation techniques include but are not limited to masking and using synthetic data. There are various factors that are decisive, but much depends on whether a technical or a legal approach is adopted. Also, in technical literature, various types of anonymity, each with their own emphasis on different factors, have been put forward, most importantly: *k*-anonymity, *l*-diversity, *t*-closeness and ε-differential privacy.

For aggregation, a difference can be made between, inter alia, aggregation based on third parties, aggregation based on data perturbation, and aggregation based on cryptography. Each of those underlines different factors that are deemed to be decisive. Perhaps the most important technique in terms of aggregating data, especially in light of data disclosure, is SDC. There is no fixed standard for SDC; each agency may adopt its own factors, standards and thresholds, taking account of the dataset, its value and potential privacy risks.

Several pseudonymisation techniques exist, most importantly, for the purposes of this study: hashing, key hashing, salt hashing and pepper hashing. Legally, encryption is seen as a sub-set of pseudonymisation. Several encryption techniques exist, most importantly: symmetric encryption, asymmetric encryption, homomorphic encryption and multiparty computation, which is more than merely an encryption technique; it is a technique that deals with protocols that allow a set of parties to jointly compute a function of their inputs or identifiers while avoiding revealing anything but the output of said function.

4. What technical developments in the area of anonymisation and pseudonymisation of personal data are to be expected in the coming years?

Most experts interviewed and the literature evaluated for this study do not expect a technological revolution in terms of anonymisation and pseudonymisation, but rather expect the game of cat and mouse to continue over the coming years. However, due to the general availability of data and the general availability of technology, it may become even harder to arrive at anonymous or pseudonymous data. Quantum Computing, as mentioned, could have an important impact on encryption. In addition, Deep Learning is a technology that is expected to gain even more prominence in the coming years. Both technologies could have a detrimental effect on privacy, but they could also be put to its advantage. Post-quantum encryption is believed to be much more safe than current forms of encryption, and deep privacy tools (privacy tools based in deep learning models) are currently being developed.

5. What can be said, from a legal and technical perspective, about the interpretation of the concept of 'means reasonably likely to be used'? What means can be considered reasonably likely to be used and what factors play a role in this?

From a legal perspective, both the CJEU and the WP29 have emphasised time and again that the assessment of what means are deemed to be reasonably likely to be used should be done on a case-by-case basis, taking into account all relevant circumstances of the case and having an eye to various relevant factors that are not determinative in themselves, such as the costs of and the amount of time required for identification, the available technology at the time of the processing and technological developments. Although these are objective criteria in and of themselves, their interpretation depends on context. Thus, although the distinction between non-personal and personal is binary and absolute in its legal effect, the criteria to determine whether data are anonymous are highly contextual.

From a technical perspective, the contextual approach is most apparent. Most technical experts do not believe in absolute or full anonymity, but rather point to a scale of how difficult it is to de-anonymise or re-identify a database. Since technological capabilities for de-anonymisation are evolving, an assessment of the technical standards to anonymise data might need to be permanent or periodical. In view of this fact, a black-and-white distinction between anonymous and non-anonymous data is not obvious; rather, from a technical perspective, it might be more appropriate to work with a scale under which the more anonymous data is, the less (strict) data protection standards apply. There is no exhaustive list of factors from a technological perspective that should be taken into account in order to determine the means reasonably likely (a legal notion that is not standardised in most technological discourse).

6. How does the answer to question 5 relate to developments in current and expected techniques to achieve anonymisation and pseudonymisation?

The general availability of open data and the general availability of data technologies will have a threefold impact on the possibilities of achieving anonymisation and pseudonymisation.

First, the nature of the data in Big Data processes is not stable, but volatile. A dataset containing ordinary personal data can be linked to and enriched with another dataset to derive sensitive data; the data can then be aggregated or stripped of identifiers and become non-personal, such as aggregate or anonymous data; subsequently, the data can be de-anonymised or integrated into another dataset in order to create personal data again. All this can happen in a split second. The question is, therefore, whether it makes sense to work with well-defined categories if the same datum or dataset can fall into a different category from one second to the next and into still another the next second.

Second, as a consequence of the previous issue, it is increasingly difficult to determine the status of data precisely. In order to determine the current status of a datum or dataset, the expected future status of the data must be taken into account. Given the general availability of technologies and the minimal investment required, it is increasingly likely that if a database is shared or otherwise made available, there will be a party who will combine it with other data, enrich it with data scraped from the internet or merge it into an existing dataset, but also that there are other parties who will not. The legal category to which the data belongs is therefore no longer a quality of the data itself, but a product of a data controller's efforts and investments. Consequently, it is arguable whether anonymisation or pseudonymisation can be achieved in a context where the determination of the status of data is hardly attainable.

Third, modern data processing operations are increasingly based on aggregate data, which can also have very large individual and social consequences. Profiling target groups rather than individuals is becoming a prevalent processing operation in the information society. The consequences of these

activities can be negative for a group, without the damage being directly relatable to individuals. The idea that the more sensitive the data are and the more directly they can be linked to a person, the more strictly its processing should be regulated can therefore be questioned. In addition, there is also the question of whether the focus on the identifiability of an individual (natural person) and, subsequently, the notions of anonymisation and pseudonymisation which are built thereon, are still viable in the 21st century.

7. When is it reasonable to say that data can no longer be linked back to an individual and that the dataset of which they are part can be considered anonymous?

While from a legal perspective, there is a difference between non-personal and personal data, from a technical perspective, this distinction falls apart into at least three relevant subcategories:

1. The situation in which data was never personal before, but might be, such as when weather data are used to make decisions about individual farmers' insurance.
2. The situation in which data were personal, but the identifiers have been stripped or data has been rendered anonymous in such a manner that the data subject cannot be identified nor made identifiable. Here, the danger is that data are re-identified or de-anonymised.
3. The situation in which data are aggregated. Here, the danger exists that data can be de-aggregated, that two datasets combined can yield personal data and that aggregate data can be used to making decisions that have an impact on individual data subjects or single them out, without knowing their identity.

Different threats are posed by each of those scenarios. From the technological domain, it is clear that it is almost never reasonable to state that data can no longer be linked back to an individual. There are always risks of de-anonymisation, there are always possibilities of data composition and one can never exclude the possibility that data will be used for singling out non-identified individuals or for developing decision trees that have an impact on groups and/or individuals. As a result, it is increasingly difficult to affirm that data can no longer be linked back to an individual and that the dataset of which they are part can be considered anonymous.

8. To what extent is the test for indirect identifiability objectifiable?

Few cues have been found to make the test more objectifiable. It is important to underline that making the test objective was not the desire of the EU regulator. On the contrary, the current open, contextual and fluid set of criteria were favoured over the more restrictive ones that were considered and rejected. For example, the initial proposal for the Data Protection Directive did not contain the notion of anonymity, but rather that of 'depersonalisation', which was understood as modifying information in such a way that it could no longer be associated with a specific individual. The explanatory memorandum provided that "[a]n item of data can be regarded as depersonalized even if it could theoretically be repersonalized with the help of disproportionate technical and financial resources". At the same time, the explanatory memorandum defined depersonalisation as "modify[ing] personal data in such a way that the information they contain can no longer be associated with a specific individual or an individual capable of being determined except at the price of an excessive effort." Excessive effort is still contextual, but less so than "all means reasonably likely"; the threshold is also clearly different.

Few cues have been found in this study for making the test of indirect identifiability more objective other than removing the notion of 'identifiability', which was not part of the definition of personal data under the data protection regimes from before 1995 or the list of factors to be included for determining what means should be deemed reasonably likely to be used. Perhaps the only concrete suggestion that was identified is putting a time limitation or a horizon to the evaluation of the means reasonably likely to be used. It is almost always highly likely that, in 20 years' time, data that are anonymous now can be de-anonymised. Under the current legal regime, when data are stored for that long, such means

reasonably likely to be used must be taken into account when determining whether the data protection regime applies, while it is next to impossible to foresee how the technological landscape and the availability of data will evolve in the next 20 years.

9. To what extent and in what cases can there be under-regulation when data are no longer linked to individuals through anonymisation and therefore do not fall within the scope of the GDPR?
10. To what extent and in what cases can there be overregulation when more and more data can be easily linked to individuals through new techniques, undoing measures of anonymisation and pseudonymisation?

Answering questions 9 and 10 depends on what is deemed to be the regulatory objective of the data protection regime: is the data protection framework to be considered from a protective angle or from the perspective of facilitating data processing within a set framework, or as a combination between both? Is the protective rationale to be understood as primarily providing protection to individual interests or to group and societal interests? Should the data protection regime be understood as laying down limitations for data processing or as providing a framework for using and sharing data? Is the protective rationale best served by limitations, or can more data processing sometimes be required to serve the best interests of individuals and/or society? Is the rationale of facilitating data use best served by an open and contextual framework or by setting strict and clear rules within which data processing is deemed legitimate? This study has not been able to give a determinative answer to these questions, but has indicated that dependent on these answers, different regulatory gaps and dangers for over-regulation and/or under-regulation will be found.

For example, whether there is under-regulation because 'personal data' is linked only to the identifiability of natural persons and because the data protection framework refers primarily to the interests of the data subject depends on what rationale the data protection framework is said to protect. If it is considered that the data protection framework is or should be providing protection to more general, group or societal interests, then certainly, there may be a matter of under-regulation due to the fact that processing aggregate and anonymous data is not covered under the current regime. Likewise, whether the trend of courts and advisory bodies to expand the scope of personal data and the material scope of the data protection framework leads to overregulation is dependent on whether the emphasis is placed on the protective rationale of the data protection framework, in which case there would be no over-regulation, but on the contrary, this approach could be deemed laudable, or on the rationale facilitating data processing, it may be deemed stifling.

11. How will the current and future technical developments affect the GDPR and the legal protection of data in a broad sense in the coming period?

It is clear that technological developments and the general availability of data, now and in the future, will mean anonymisation will become increasingly difficult. The status of data will become increasingly volatile, and this will be due less and less to the nature of the data and datasets themselves and more and more due to data controllers' efforts. Legal categories will become more and more fluid and porous, and one database may be legally qualified differently depending on the party that has access to it. A database that only contains non-personal data in isolation may be turned into personal data by being combined, then used to infer sensitive personal data, only to be aggregated and anonymised the next moment again. Given these trends and given the notions of 'identifiability' and 'all the means reasonably likely to be used', more and more data, if not all, will fall under the data protection framework.

This study did not find different scenarios for how the technological realm and the availability of open data will develop over time – literature, interviewed experts and experts invited to the workshop held for this study all point in the same direction. Several scenarios were found, however, for how the legal

regime could respond to the increased availability of open data and the general availability of technology. Five strategies were deduced from the suggestions: leaving the current data protection framework intact, focussing on clearer data categories, focussing more on contextuality, using different data categories and regulatory regimes attached to them, or focussing on a full-blown contextual data protection framework.

Nederlandse samenvatting

Introductie

De Algemene Verordening Gegevensbescherming (AVG) is misschien wel het belangrijkste kader voor het digitale domein in Europa en daarbuiten. De AVG stelt regels aan- en bevat normen voor de verwerking van gegevens, legt verplichtingen vast voor personen en organisaties die gegevens verwerken (verwerkingsverantwoordelijken) en kent rechten toe aan personen van wie gegevens worden verwerkt (betrokkenen). Hoewel pas in 2016 aangenomen, stammen de regels in essentie uit de jaren 70 van de vorige eeuw. Doorslaggevend voor de toepassing van het gegevensbeschermingskader was toen, en is vandaag de dag nog steeds, of de gegevens die worden verwerkt informatie van een identificeerbaar individu (natuurlijke persoon) betreffen.

Hoewel een dergelijke vaststelling in de jaren 70 van de vorige eeuw relatief eenvoudig was, is die in de loop van de tijd steeds complexer geworden, vooral in het licht van technologische ontwikkelingen, de algemene toegankelijkheid van technologieën en het streven naar meer open data. Deze ontwikkelingen hebben tot gevolg dat het steeds makkelijker is om persoonsgegevens af te leiden uit datasets die dergelijke gegevens op het eerste gezicht niet lijken te bevatten. Ze hebben ook tot gevolg dat de juridische status van data steeds meer fluïde wordt: doordat data worden gedeeld tussen partijen en de verwerkingen van datasets aanzienlijk verschillen, kan dezelfde dataset het ene moment worden gekwalificeerd als persoonsgegevens en het andere moment niet, of als persoonsgegevens in handen van partij A maar tegelijkertijd als geen persoonsgegevens in handen van partij B.

Daarom is in de loop der tijd in het wettelijke kader het begrip persoonsgegevens uitgebreid. Met name in 1995 breidde de voorloper van de Algemene Verordening Gegevensbescherming, de Richtlijn Gegevensbescherming, de reikwijdte van dit begrip aanzienlijk uit en daarmee ook het aantal datasets dat onder het bereik van het gegevensbeschermingsregime viel. Bij persoonsgegevens gaat het niet alleen om directe maar ook om indirecte informatie, dat wil zeggen gegevens, zoals beschrijvingen, waaruit de identiteit van een persoon kan worden afgeleid. Bij persoonsgegevens gaat het niet alleen om identificerende gegevens, dat wil zeggen gegevens die op dit moment tot een bepaalde persoon kunnen leiden, maar ook identificeerbare gegevens, of met andere woorden gegevens die op dit moment niet tot een bepaalde persoon leiden, maar in de toekomst mogelijk wel. Om te bepalen of een dataset identificeerbare gegevens bevat, moet rekening worden gehouden met alle middelen waarvan redelijkerwijs valt te verwachten dat zij worden gebruikt om gegevens aan een persoon te koppelen. Ten slotte is het niet nodig om de identiteit van een persoon te kennen; als gegevens worden gebruikt om een beslissing te nemen over een specifieke persoon wiens identiteit onbekend is, is het gegevensbeschermingsregime ook van toepassing.

Deze wetswijzigingen hebben geleid tot een substantiële uitbreiding van het toepassingsgebied van het gegevensbeschermingsregime. Tegelijkertijd blijft het begrip persoonsgegevens de bepalende factor bij de beslissing of de gegevensbeschermingsregels van toepassing zijn. In tegenstelling tot het restrictieve regime voor de verwerking van persoonsgegevens heeft de Europese Unie (EU) een ander kader vastgesteld voor de verwerking van niet-persoonsgegevens. De Verordening betreffende het vrije verkeer van niet-persoonsgegevens houdt in wezen in dat er geen beperkingen mogen worden gesteld, noch door de publieke sector, noch door de private sector, met betrekking tot het vrije verkeer van niet-persoonsgegevens. De juridische kwalificatie of een dataset al dan niet persoonsgegevens bevat betekent dus dat er een reguleringskader van bijna 180 graden verschil van toepassing is (hoewel de voorgestelde ‘Data Governance Act’ de zaken nog ingewikkelder kan maken).

Er zijn ook belangrijke technologische en maatschappelijke ontwikkelingen. Big Data, Kunstmatige Intelligentie, Quantum Computing en andere technieken maken het nog gemakkelijker om persoonsgegevens af te leiden uit geaggregeerde, geanonimiseerde of versleutelde datasets; de algemene

toegankelijkheid van technologieën maakt het nog moeilijker om de toekomstige status van een dataset te bepalen; en het voortdurende streven naar open data en het hergebruik van overheidsinformatie betekent dat de juridische status van data nog meer fluïde zal worden. In het licht van deze nieuwe uitdagingen is het de vraag hoe het juridische regime hierop moet reageren. Moet het begrip persoonsgegevens verder worden opgerekt? Zo ja, zou dat in de praktijk niet betekenen dat alle gegevens als persoonsgegevens worden aangemerkt? Moet het huidige onderscheid tussen persoonsgegevens en niet-persoonsgegevens behouden blijven, of moet er een restrictiever regime komen voor niet-persoonsgegevens? En wat betekenen deze ontwikkelingen voor andere gegevenscategorieën in de Algemene Verordening Gegevensbescherming, zoals pseudonieme gegevens en gevoelige (bijzondere) persoonsgegevens?

Tegen deze achtergrond is de onderzoeksvraag voor dit onderzoek: *Welk effect hebben huidige en toekomstige technische ontwikkelingen op het gebied van anonimisering, pseudonimisering, aggregatie en identificatie van gegevens, op het gegevensbeschermingskader en de bescherming van de verschillende soorten gegevens?*

De deelvragen, die helpen bij het beantwoorden van deze onderzoeksvraag, zijn:

Identificeerbaarheid van gegevens

1. Welke (technische) middelen zijn er om (anonieme) data terug te koppelen aan individuen, en in hoeverre speelt de beschikbaarheid van andere (bijvoorbeeld open source) data een rol?
2. Welke (technische) ontwikkelingen worden de komende jaren verwacht met betrekking tot de middelen om gegevens (al dan niet opzettelijk) terug te koppelen aan personen?

Anonimisering en pseudonimisering van gegevens

3. Welke huidige en voorzienbare technische ontwikkelingen kunnen worden gebruikt voor het anonimiseren of pseudonimiseren van persoonsgegevens en welke factoren zijn daarbij bepalend?
4. Welke technische ontwikkelingen op het gebied van anonimisering en pseudonimisering van persoonsgegevens zijn de komende jaren te verwachten?

Identificeerbaarheid in relatie tot anonimisering en pseudonimisering

5. Wat kan er vanuit een juridisch en technisch perspectief worden gezegd over de invulling van het begrip ‘alle middelen waarvan redelijkerwijs valt te verwachten dat zij worden gebruikt’? Welke middelen kunnen redelijkerwijs worden gebruikt en welke factoren spelen daarbij een rol?
6. Hoe verhoudt het antwoord op vraag 5 zich tot ontwikkelingen in huidige en toekomstige anonimiserings- en pseudonimiseringstechnieken?
7. Wanneer is het redelijk om te zeggen dat gegevens niet meer terug te koppelen zijn aan een persoon en dat de dataset waarvan ze deel uitmaken als anoniem kan worden beschouwd?
8. In hoeverre is de test op indirecte identificeerbaarheid objectiveerbaar?

Gevolgen van identificeerbaarheid en anonimisering en pseudonimisering

9. In hoeverre en in welke gevallen kan er sprake zijn van onderregulering wanneer gegevens door anonimisering niet meer aan personen kunnen worden gekoppeld en dus niet binnen de reikwijdte van de Algemene Verordening Gegevensbescherming vallen?
10. In welke mate en in welke gevallen kan er sprake zijn van overregulering wanneer steeds meer gegevens eenvoudig aan individuen kunnen worden gekoppeld door middel van nieuwe technieken (het ongedaan maken van anonimisering en pseudonimisering)?

Overkoepelende analyse

11. Hoe zullen de huidige en toekomstige technische ontwikkelingen de komende periode van invloed zijn op de AVG en rechtsbescherming in brede zin?

Voor het beantwoorden van deelvragen 1-8 zijn verschillende aspecten relevant:

- de verschillende juridische begrippen en de criteria in de definities en afbakeningen;
- de beschikbaarheid van (open access) data en van dataverwerkingstechnologieën; in dit opzicht

is het streven van de Europese Unie naar open data en hergebruik van (overheids)data relevant;
- de huidige en toekomstige technologische middelen voor het anonimiseren en de-anonimiseren, aggregeren en de-aggregeren, pseudonimiseren en de-pseudonimiseren van gegevens; en
- de impact van de ontwikkelende technologische mogelijkheden en het uitbreidende datalandschap op de houdbaarheid van huidige juridische concepten en afbakeningen.

Om deelvragen 9-11 te beantwoorden, zijn verschillende aspecten relevant:
- de reguleringsdoelstelling van het gegevensbeschermingskader en daarmee het licht waarin het gevaar van zowel onder- als overregulering beoordeeld moet worden;
- de lacunes in de regelgeving die voortkomen uit de kloof tussen het juridische en het technologische domein; en
- de alternatieven voor het huidige wettelijke kader die uit voorgangers van de Europese wetgeving en uit wetsvoorstellen, literatuur en interviews kunnen worden gedistilleerd.

Bij de beantwoording van de vragen 9-11, en om te bepalen of er sprake is van onder- en/of overregulering, moet worden bepaald wat het reguleringsdoeleinde van de AVG is en zou moeten zijn. Daarbij moeten twee zaken worden onderzocht. Enerzijds is het de vraag of het gegevensbeschermingsrecht als enige of belangrijkste doel heeft om natuurlijke personen te beschermen. Verschillende auteurs wijzen erop dat de wetgeving inzake gegevensbescherming, althans aanvankelijk, vooral gericht was op de bescherming van objectieve rechtsbeginselen en algemene belangen. Anderzijds wordt in de juridische literatuur bediscussieerd in hoeverre de bescherming van natuurlijke personen de beste basis is voor toekomstige regelgeving en of deze bescherming niet moet worden uitgebreid naar groepen of de samenleving als geheel.

Voor dit onderzoek worden drie methoden ingezet:

1. Doctrinaire en juridische analyse: vier juridische onderscheiden tussen gegevens staan centraal in dit onderzoek, namelijk het onderscheid tussen: anonieme gegevens en persoonsgegevens, geaggregeerde of statistische gegevens en persoonsgegevens, pseudonieme en niet-pseudonieme persoonsgegevens en niet-gevoelige en gevoelige persoonsgegevens. Hiervoor worden de wetten van de EU en de Raad van Europa (RvE), hun wetsgeschiedenis en juridische interpretatie bestudeerd.
2. Literatuuroverzicht.
    a. Beschrijvende literatuur: technische literatuur over (de-)identificatietechnologieën en privacy/gegevensbescherming verbeterende technieken wordt bekeken.
    b. Normatieve literatuur: juridische en reguleringsliteratuur wordt bestudeerd die de uitdagingen van elke categorie gegevens beschrijft en/of nieuwe definities, perspectieven of benaderingen voor de verschillende soorten gegevens voorstelt.
3. Kwalitatieve onderzoeksmethoden.
    a. Interviews: er zijn interviews gehouden met experts met verschillende achtergronden en expertisegebieden.
    b. Workshop: aan het begin van dit onderzoek is een workshop gehouden om problemen en mismatches tussen het juridische en beleidsdomein enerzijds en de technische en praktische realiteit anderzijds te identificeren.

Het onderzoek liep langs de volgende lijnen.

Het wettelijk regime is beoordeeld op drie punten:
(1) Het huidige wettelijke regime en de bestaande definities en uitleg daarvan in literatuur of gezaghebbende adviezen zijn geanalyseerd om te bepalen hoe het bestaande wetgevend kader gegevensverwerking beoordeelt.
(2) De geschiedenis van het juridisch regime vanuit het oogpunt van de definities is om drie

redenen geëvalueerd. Ten eerste laat het zien hoe het kader voor gegevensbescherming in de loop van de tijd is gewijzigd in reactie op maatschappelijke en technologische veranderingen. Ten tweede geeft het inzicht in de logica en beweegredenen achter de huidige definities en categorisering: waarom zijn de definities zoals ze zijn en welk doel wordt nagestreefd. Meer in het algemeen werd aandacht besteed aan de discussie over de achterliggende gedachte van het gegevensbeschermingskader, aangezien dit relevant is met het oog op mogelijke toekomstige wijzigingen in het gegevensbeschermingskader. Ten derde kunnen door de verschillende definities en afbakeningen van de gegevenscategorieën en vooral de variaties die in de wetsgeschiedenis zijn besproken en overwogen, maar werden verworpen, alternatieve manieren worden gevonden om de regulering van gegevens aan te pakken.

(3) De potentiële toekomst van het gegevensbeschermingskader werd beoordeeld. De in dit onderzoek besproken technologische en maatschappelijke ontwikkelingen hebben grote invloed op de invulling en effecten van het huidige regulerende kader. Daarom wordt een overzicht gegeven van de belangrijkste ideeën voor mogelijkheden om het huidige regulerende kader te wijzigen.

Het technologische domein werd beoordeeld op drie punten.

(1) Om het beeld te schetsen van een veld dat voortdurend in beweging is, is een kort overzicht gegeven van de technologische ontwikkelingen na de Tweede Wereldoorlog. Deze beschrijving geeft de achtergrond waartegen het wettelijk kader in de loop van de tijd is gewijzigd.

(2) Het onderzoek heeft de huidige technologieën beoordeeld, met name in het licht van de verschillende juridische gegevenscategorieën en de grenzen daartussen. Deze beschrijving laat zien dat het steeds beter mogelijk wordt een dataset te de-anonimiseren en (gevoelige) persoonsgegevens af te leiden uit één of meer geaggregeerde datasets.

(3) Het onderzoek beschrijft technologische ontwikkelingen die het landschap in de toekomst mogelijk nog verder zullen veranderen. Hieruit blijkt dat de scheidslijnen tussen de verschillende juridische gegevenscategorieën zo mogelijk nog meer zullen vervagen.

Ook is er aandacht besteed aan twee maatschappelijke ontwikkelingen (hoewel deze zowel door juridische als technologische ontwikkelingen zijn ingegeven):

(1) De studie beschrijft hoe technologieën in de loop van de tijd algemeen beschikbaar zijn geworden. Hierdoor beschikken steeds meer overheidsorganisaties, bedrijven en zelfs burgers over zeer geavanceerde technologische middelen. Het gevolg van deze trend is dat als data tussen verschillende partijen worden gedeeld of openbaar worden gemaakt, het steeds waarschijnlijker wordt dat er een partij is die de juridische status van de dataset verandert.

(2) Het onderzoek verwijst kort naar de juridische en maatschappelijke druk om gegevens openbaar te maken. Dit betreft voornamelijk statistische gegevens, overheidsinformatie en niet-persoonsgegevens. Meestal zullen deze datasets op zichzelf geen persoonsgegevens bevatten, maar in combinatie met andere datasets kunnen ze worden gebruikt om (gevoelige) persoonsgegevens te genereren. Bovendien is het, gezien de vooruitgang en de algemene toegankelijkheid van technologieën, steeds waarschijnlijker dat er een partij zal zijn die voldoende middelen zal investeren om een dataset te de-anonimiseren of opnieuw te identificeren.

Dit onderzoek buigt zich over vier juridische gegevenscategorieën die zijn verankerd in de Algemene Verordening Gegevensbescherming, naast persoonsgegevens zijn er: geanonimiseerde gegevens, geaggregeerde of statistische gegevens, gepseudonimiseerde persoonsgegevens en gevoelige persoonsgegevens. Hieronder wordt een samenvatting gegeven van de belangrijkste bevindingen op de volgende punten: (1) de huidige regulering van de verschillende gegevenscategorieën; (2) de twee, soms tegenstrijdige, benaderingen van gegevensregulering die door het kader voor gegevensbescherming lopen; (3) de algemene toegankelijkheid van technologieën en het streven naar open data en het hergebruik van overheidsinformatie; (4) de impact van het veranderende technologische landschap op

de regulering van data; (5) de lacunes die bestaan tussen het huidige reguleringsstelsel en de veranderende technologische realiteit; (6) de alternatieven voor het huidige reguleringsregime die in de literatuur en elders worden gesuggereerd om deze lacunes te dichten; (7) de overkoepelende reguleringsdoelstelling van het gegevensbeschermingskader in het licht waarvan mogelijke wijzigingen moeten worden beoordeeld; (8) de gevaren van over- en onderregulering veroorzaakt door de mismatch tussen het juridische en het technologische domein; (9) en de mogelijke manieren om de bestaande hiaten tussen de twee domeinen op te lossen. Tot slot (10) worden de onderzoeksvraag en deelvragen beantwoord.

1. Juridische categorieën en de elementen daarvan

Dit onderzoek heeft zich gericht op vier gegevenscategorieën onder het gegevensbeschermingsregime. Naast persoonsgegevens beschouwt het onderzoek anonieme gegevens, geaggregeerde of statistische gegevens, pseudonieme persoonsgegevens en gevoelige persoonsgegevens.

Anonieme gegevens

In dit onderzoek is de scheidsgrens tussen persoonsgegevens en anonieme gegevens onderzocht. Anonimiseren betekent het wegnemen van direct of indirect geïdentificeerde of identificeerbare data in data. Als gegevens correct geanonimiseerd zijn is de AVG niet, maar de Verordening vrij verkeer van niet-persoonsgegevens wel van toepassing. Uit de formele definitie van persoonsgegevens (artikel 4 lid 1 AVG), de relevante overwegingen (14, 26, 27 en 30), en de interpretatie door het Hof van Justitie van de Europese Unie (HvJ EU) en de Artikel 29 Werkgroep, vallen minstens vier punten op te maken:

1. Niet alleen direct identificerende gegevens maar ook indirect identificerende gegevens, en niet alleen identificerende gegevens maar ook identificeerbare gegevens, moeten als persoonsgegevens worden aangemerkt. Dat laatste betekent dat de huidige status van gegevens niet bepalend is; om de juridische categorisering te bepalen (zijn gegevens juridisch gezien persoonsgegevens of niet?), moet rekening worden gehouden met de waarschijnlijke toekomstige status ervan. Zoals de Artikel 29 Werkgroep heeft benadrukt moet de verwerkingsverantwoordelijke rekening houden met de mogelijkheid van identificatie die zich ook over 9 jaar kan voordoen. Dit heeft grote gevolgen, met name voor open data, die permanent online blijft en door verschillende partijen zal worden gebruikt.
2. Om vast te stellen of gegevens persoonsgegevens zijn, dient rekening te worden gehouden met alle middelen die redelijkerwijs voor identificatie kunnen worden gebruikt. Om de waarschijnlijkheid van identificatie vast te stellen, moet worden gekeken naar de kosten en de hoeveelheid tijd die nodig is voor identificatie, de beschikbare technologie op het moment van de verwerking, en toekomstige technologische ontwikkelingen. Hoewel dit op zichzelf objectief verifieerbare criteria zijn, hangt de interpretatie ervan, zoals zowel de Artikel 29 Werkgroep als het HvJ EU keer op keer hebben benadrukt, af van de context.
3. De vraag is niet alleen of de verwerkingsverantwoordelijke zelf nu of in de toekomst persoonsgegevens kan verkrijgen uit een dataset, maar ook of een partij die toegang heeft tot de gegevens dat kan. Dit is wederom met name van belang wanneer gegevens online beschikbaar worden gesteld of met meerdere partijen worden gedeeld. Hoe meer partijen toegang hebben tot een database, hoe groter de kans dat iemand persoonsgegevens uit de dataset afleidt, terwijl tegelijkertijd de middelen om te verifiëren of iemand dit heeft gedaan, wanneer en waarom, afnemen.
4. Identificatie is niet vereist; het kunnen uitlichten van een persoon is voldoende. Als een internetbedrijf niet weet wie een persoon is, maar wel gepersonaliseerde advertenties kan tonen aan account 87&^%11!, dan is dat in principe voldoende om de gegevens juridisch te kwalificeren als persoonsgegevens. Evenzo is dit het geval wanneer een verzekeraar aanvragen afwijst van een persoon (zonder zijn naam te kennen) uit een gebied met een specifieke postcode. Op een vergelijkbare manier heeft de Artikel 29 Werkgroep benadrukt dat gegevens kunnen

worden beschouwd als "betrekking hebbend op" een persoon, omdat het gebruik ervan waarschijnlijk gevolgen zal hebben voor de rechten en belangen van een bepaalde persoon, rekening houdend met alle omstandigheden van het concrete geval. De Artikel 29 Werkgroep benadrukte dat het niet nodig is dat het potentiële resultaat een grote impact heeft.

Geaggregeerde gegevens

Door aggregatie kunnen gegevens geanonimiseerd worden door de gegevens niet langer te behandelen op het niveau van n = 1, maar op het niveau van n = 20, n = 100, enz. De analyse van geaggregeerde gegevens kan leiden tot informatie zoals - in zeer basale termen - van de 100.000 mensen met een groene auto heeft 34% een witte bank in de woonkamer. Deze gegevens worden in principe niet als persoonsgegevens beschouwd. Wanneer partijen echter op basis van geaggregeerde gegevens handelen op een manier die directe gevolgen heeft voor natuurlijke personen, dan kan dat wel, bijvoorbeeld wanneer een autobedrijf advertenties voor witte banken stuurt naar alle mensen die een groene auto hebben gekocht. Door de hele AVG heen zijn er verwijzingen naar statistische gegevens en geaggregeerde gegevens bedoeld voor onderzoeksdoeleinden. De AVG onderschrijft de publieke belangen die met statistische analyse gediend kunnen worden (statistische analyse door het Centraal Bureau voor de Statistiek is bijvoorbeeld essentieel voor op informatie gebaseerde beleidsvorming door de overheid). Als gegevens zodanig worden geaggregeerd dat er geen individuele gegevens kunnen worden geëxtraheerd noch worden gebruikt op een manier die directe gevolgen heeft voor concrete personen, is de AVG niet van toepassing. In dat geval kunnen de regels voor het verwerken van statistische gegevens gelden, die normen inhouden voor onder meer vertrouwelijkheid en veiligheid.

Wanneer persoonsgegevens worden gebruikt voor statistische verwerkingen, is de AVG van toepassing, maar laat deze ruimte voor uitzonderingen op nationaal niveau. Artikel 85 AVG maakt uitzonderingen mogelijk voor de verwerking van persoonsgegevens in het kader van de vrijheid van meningsuiting; artikel 86 AVG bepaalt dat persoonsgegevens in officiële documenten die in het bezit zijn van een (semi) publieke instelling, door die instelling mogen worden bekendgemaakt in overeenstemming met wet- en regelgeving om de toegang van het publiek tot officiële documenten in overeenstemming te brengen met het recht op bescherming van persoonsgegevens; en artikel 89 AVG bepaalt dat de lidstaten vrijstellingen kunnen aannemen, met name ten aanzien van de rechten van betrokkenen, wanneer persoonsgegevens worden verwerkt voor archiveringsdoeleinden in het algemeen belang, wetenschappelijke of historische onderzoeksdoeleinden of statistische doeleinden. Er zijn geen factoren uiteengezet om te bepalen wanneer een database zodanig wordt geaggregeerd dat deze kwalificeert als niet-persoonsgegeven. Dit is afhankelijk van de omstandigheden van het geval, rekening houdend met de eerder besproken algemene elementen.

Pseudonieme gegevens

De AVG is van toepassing op gepseudonimiseerde gegevens, maar er zijn enkele uitzonderingen van toepassing op de verplichtingen van verwerkingsverantwoordelijken wanneer zij gegevens hebben gepseudonimiseerd. Bovendien wordt pseudonimisering beschouwd als een manier om technische en organisatorische veiligheidsstandaarden te implementeren, specifieke verplichtingen die zijn vastgelegd in het kader voor gegevensbescherming. De AVG (Artikel 4 lid 5) definieert 'pseudonimisering' als het verwerken van persoonsgegevens op een zodanige manier dat de persoonsgegevens niet meer aan een specifieke betrokkene kunnen worden toegeschreven zonder het gebruik van aanvullende informatie, op voorwaarde dat dergelijke aanvullende informatie afzonderlijk wordt bewaard en is onderworpen aan technische en organisatorische maatregelen om ervoor te zorgen dat de persoonsgegevens niet worden toegeschreven aan een geïdentificeerde of identificeerbare natuurlijke persoon. Overweging 28 maakt duidelijk dat pseudonimisering van persoonsgegevens de risico's voor de betrokkenen kan verkleinen. Daarom wordt, zoals uiteengezet in overweging 29, pseudonimisering gestimuleerd door de AVG.

Het begrip pseudonimisering is nieuw in de AVG; het speelde geen rol in eerdere gegevensbeschermingsregelingen. Hoewel de Verordening benadrukt dat andere technieken voor een veilige verwerking van persoonsgegevens niet worden uitgesloten door het feit dat pseudonieme gegevens apart worden gedefinieerd, wordt aan deze techniek wel een bijzondere status toegekend. Wat de juiste interpretatie van dit juridische begrip complex maakt is dat de AVG vaak pseudonimisering in één adem noemt met encryptie, een term die niet apart wordt gedefinieerd. De Artikel 29 Werkgroep steunt het toegenomen gebruik van pseudonimiseringstechnieken, waarvan zij vijf belangrijke technieken onderscheidt, waaronder encryptie, en beschouwt (bepaalde vormen van) encryptie als een subset van pseudonimiseringstechnieken.

Gevoelige persoonsgegevens

Gevoelige persoonsgegevens worden onder het gegevensbeschermingsregime apart gedefinieerd ten opzichte van 'gewone' persoonsgegevens. Bijzondere of gevoelige persoonsgegevens zijn duidelijk omschreven en afgebakend; de verwerking geldt per definitie als potentieel schadelijk voor de belangen van natuurlijke personen. Het verwerken van gevoelige gegevens is in principe verboden, al geldt er een groot aantal uitzonderingen op dat verbod. Gevoelige gegevens worden gedefinieerd als persoonsgegevens waaruit ras of etnische afkomst, politieke opvattingen, religieuze of levensbeschouwelijke overtuigingen of lidmaatschap van een vakbond blijken, en de verwerking van genetische gegevens, biometrische gegevens met het oog op de unieke identificatie van een natuurlijke persoon, gegevens over gezondheid of gegevens over het seksleven of de seksuele geaardheid van een natuurlijk persoon. De verwerking van strafrechtelijke gegevens door rechtshandhavingsinstanties valt onder de zogenoemde Politierichtlijn.

Het HvJ EU heeft een ruime interpretatie gegeven aan wat als gevoelige persoonsgegevens moet worden beschouwd. In de *Lindqvist*-zaak had iemand bijvoorbeeld op een blog geschreven dat een collega vanwege medische redenen deeltijd werkte omdat ze een voetblessure had opgelopen. De vraag of informatie over het hebben van een voetblessure reeds kwalificeert als 'medische gegevens' werd door de rechter slechts kort, staccato en bevestigend beantwoord. Een andere zaak, die van *V.*, betrof de overdracht van een medisch dossier in het kader van de arbeidsverhouding. De rechter wees erop dat medische gegevens bijzonder gevoelige gegevens zijn, waardoor schijnbaar een hiërarchie ontstaat tussen verschillende categorieën gevoelige persoonsgegevens en medische gegevens bovenaan komen te staan.

2. Juridisch regime: de categorale en de contextuele benadering

Er is een spanning tussen twee reguleringsbenaderingen op het gebied van gegevensbescherming: een contextuele en een categorale benadering, een benadering die rekening houdt met de omstandigheden van het geval en een benadering die is gebaseerd op vaste definities en duidelijke regels die aan de definities zijn gekoppeld. Elk van deze benaderingen heeft duidelijke voor- en nadelen. De eerste benadering kan per scenario met alle relevante aspecten rekening houden, past zich beter aan, aan veranderende omstandigheden en loopt dus niet het risico achterhaald te zijn of omzeild te worden. Fluïde en contextuele reguleringsbenaderingen hebben echter het nadeel dat ze vaag zijn en weinig rechtszekerheid bieden, zowel voor de verwerkingsverantwoordelijke als voor de betrokkene. De tweede benadering lost dit probleem op: het geeft een duidelijke reeks definities en categorieën en koppelt daaraan een duidelijke reeks regels. Maar het nadeel is ook duidelijk, namelijk het risico omzeild te worden en verouderd te raken; ook is deze benadering minder granulair dan een contextuele benadering.

Er is een diepe ambivalentie in de reguleringsbenadering op dit punt.

Op het eerste gezicht is de categorale benadering het duidelijkst. Zo had de ontkoppeling van het recht op gegevensbescherming van het recht op privacy te maken met een de-contextualisering van het recht. In het mensenrechtenkader wordt een claim beoordeeld op zowel de ratione materiae (valt de klacht onder de materiële reikwijdte van het ingeroepen artikel?) als de ratione personae (kan de verzoeker bewijzen in aanmerkelijke mate getroffen te zijn?). Wat dat tweede beginsel betreft, geldt een aanzienlijke drempel, aangezien verzoekers moeten kunnen aantonen dat zij directe, individualiseerbare en substantiële schade hebben geleden. In het kader voor gegevensbescherming worden beide principes samengevoegd. Dit betekent dat elke verwerking van persoonsgegevens, hoe alledaags en betekenisloos ook, wordt beschouwd als verwerking van persoonsgegevens, waarop de AVG van toepassing is. Bijgevolg wordt het contextuele of op schade gebaseerde element dat essentieel is voor evaluaties van mensenrechtenvraagstukken weggelaten uit de gegevensbeschermingsregeling. De toepassing van het gegevensbeschermingsregime, anders dan bijvoorbeeld het recht op privacy, is niet afhankelijk van de vraag of er schade is toegebracht aan een eiser of rechthebbende.

Daarnaast is het duidelijk dat het gegevensbeschermingskader werkt met een binair onderscheid tussen persoonsgegevens en niet-persoonsgegevens. De EU heeft persoonsgegevens voorzien van de hoogste vorm van rechtsbescherming ter wereld, via de AVG en de Politierichtlijn, terwijl de EU met betrekking tot de verwerking van niet-persoonsgegevens expliciet beperkingen die zijn opgelegd door de private en publieke sector organisaties ontmoedigt door middel van de Verordening over het vrije verkeer van niet-persoonsgegevens. Omdat het onderscheid tussen persoonsgegevens en niet-persoonsgegevens binair is (alhoewel de voorgestelde Data Governance Act dit beeld misschien zal compliceren), zal de vraag of een dataset als een van beide wordt gecategoriseerd, een reguleringsverschil van 180 graden betekenen. Ook ten aanzien van zowel pseudonieme gegevens als gevoelige persoonsgegevens is sprake van een binaire benadering: gegevens zijn pseudoniem of niet, persoonsgegevens zijn gevoelig of niet. Met betrekking tot het laatste type gegevens is de categorale benadering nog duidelijker. De AVG bevat een beperkte en uitputtende lijst van soorten gegevens die als gevoelig worden beschouwd. De verwerking van dergelijke gegevens is in principe verboden.

Een laatste punt dat moet worden benadrukt, is dat het kader voor gegevensbescherming als geheel gebaseerd is op binaire onderscheiden en wordt gekenmerkt door een categorale benadering. Zo worden duidelijke verschillen tussen verschillende actoren, zoals de verwerkingsverantwoordelijke, de gegevensverwerker en de betrokkene, vastgelegd. Elk van deze actoren heeft een duidelijk omschreven rol, een reeks verplichtingen, rechten en regelgevende verantwoordelijkheden. Een partij kan niet tegelijkertijd gegevensverwerker en verwerkingsverantwoordelijke zijn met betrekking tot dezelfde gegevensverwerking: het is een kwestie van of/of. Evenzo is bij een gegevensverwerking een partij ofwel een (mede)verantwoordelijke ofwel een betrokkene.

Anderzijds is een contextuele benadering zichtbaar. Hoewel het onderscheid tussen persoonsgegevens en niet-persoonsgegevens bijvoorbeeld binair is, omvat de definitie van persoonsgegevens een contextueel aspect. Het begrip ‘identificeerbaar’ houdt in dat gegevens die op dit moment geen identificatie van een persoon mogelijk maken, maar dit in de toekomst wel mogelijk zullen maken, nu al als persoonsgegevens worden aangemerkt.

Bovendien, hoewel de categorie van pseudonieme gegevens op zichzelf binair is - gegevens zijn pseudoniem of niet - wordt deze categorie door velen gezien als een tussencategorie tussen persoons- en niet-persoonsgegevens. Pseudonieme gegevens zijn niet anoniem en daarom is de AVG van toepassing, maar ze zijn niet zo eenvoudig te koppelen aan een geïdentificeerde persoon. Daarom staat de AVG een aantal uitzonderingen toe wanneer gegevens worden gepseudonimiseerd. Evenzo, hoewel het onderscheid tussen niet-gevoelige en gevoelige gegevens vaak als absoluut wordt gepresenteerd, zijn alle verschillende rechten en plichten van toepassing op zowel de verwerking van gevoelige als niet-gevoelige persoonsgegevens. Het enige verschil is de legitieme grond voor het verwerken van de gegevens (artikel 6 en artikel 9) en hoewel artikel 9 als uitgangspunt neemt dat het verwerken van

gevoelige gegevens verboden is, somt het een groot aantal uitzonderingen op dit verbod op, waardoor het verschil tussen de verwerking van gevoelige en niet-gevoelige gegevens minder binair is dan op het eerste gezicht lijkt.

Wanneer het gegevensbeschermingsregime van toepassing is, zijn de meeste verplichtingen en vereisten contextafhankelijk, wat in het algemeen betekent dat hoe meer gegevens worden verzameld, hoe gevoeliger die gegevens zijn, hoe hoger het risico van gegevensverwerking of hoe meer partijen erbij betrokken zijn, hoe strenger de regels en verplichtingen moeten worden geïnterpreteerd. Deze contextuele benadering is van toepassing op de verplichting om een gegevensbeschermingsbeleid te implementeren, technische en organisatorische beveiligingsmaatregelen te nemen en om aan 'data protection by design' of 'by default' te doen. Ook kennen de verplichtingen en vereisten uit de AVG bepaalde contextafhankelijke beperkingen en uitzonderingen. Zo geldt het documentatievereiste niet voor kleine organisaties die zich niet bezighouden met risicovolle verwerkingen; een gegevensbeschermingseffectbeoordeling hoeft alleen te worden uitgevoerd wanneer potentiële schade waarschijnlijk is; en organisaties uit de particuliere sector hoeven alleen een functionaris voor gegevensbescherming aan te stellen wanneer hun kernactiviteiten bestaan uit het regelmatig en systematisch monitoren van betrokkenen op grote schaal of ze grootschalige verwerking van gevoelige gegevens verrichten en de melding van datalekken is afhankelijk van de schade die waarschijnlijk uit de inbreuk voortvloeit. De kernregels uit het gegevensbeschermingskader zijn dan ook zeer contextueel.

Ten slotte moet worden benadrukt dat hoewel de Europese benadering van privacy en gegevensbescherming vaak in contrast wordt gebracht met de Amerikaanse (de eerste een omnibusbenadering, de tweede een sectorale benadering), het contrast minder scherp is dan vaak wordt gedacht. De EU maakt expliciet onderscheid tussen twee contexten wanneer zij gegevensbeschermingsregels toepast: de algemene context, die onder de AVG valt, en de wetshandhavingscontext waarop de Politierichtlijn van toepassing is. Daarnaast bevordert de AVG het gebruik van gedragscodes, waarmee sectoren hun eigen invulling en specificatie aan het gegevensbeschermingsregime kunnen geven. Dat deze mogelijkheid nauwelijks wordt gebruikt omdat sectoren vrezen voor de administratieve lasten van het uitvoeren van toezicht en het afhandelen van klachten, betekent niet dat dit niet wordt gestimuleerd door de AVG.

3. De impact van de beschikbaarheid van data en datatechnologieën op de wettelijke regulering van data

De beschikbaarheid van data groeit exponentieel. Sinds de jaren zeventig, toen de eerste grotere databases ontstonden, en nu, vijftig jaar later, is het datalandschap ingrijpend veranderd. Niet alleen worden er meer data verzameld en beschikbaar gesteld, maar fundamenteler, de samenleving is veranderd van een analoge naar een gedataficeerde samenleving, waarin vrijwel alle aspecten van het leven worden gevolgd met sensoren, cookies, camera's en satellieten. Niet alleen overheden gebruiken monitoringstechnieken om de verschillende aspecten van het leven te monitoren, ook grote internetbedrijven en steeds meer data-gedreven bedrijven doen dit. Ook burgers hebben toegang tot allerhande spyware, drones en andere sensorische producten om gegevens over zichzelf en anderen te verzamelen. Deze gegevens worden gedeeld via intermediaire platforms, opgeslagen in de 'cloud' en beschikbaar gesteld op besloten of open platforms. Een andere trend is dat, met Web 2.0, door gebruikers gegenereerde inhoud (sociale netwerken) is geëxplodeerd, en daarom zijn gebruikers zelf een belangrijke bron van persoonsgegevens (van zichzelf en hun vrienden) geworden.

Er is ook een juridisch streven om gegevens vrij te geven. In de Europese Unie zijn er verschillende wetten die partijen verplichten zich open te stellen. Zo stelt de Open Access-richtlijn voor dat de lidstaten zoveel mogelijk overheidsinformatie gratis, in open access en herbruikbaar formaat openbaar maken. De richtlijn vennootschapsrecht verplicht de lidstaten om de nodige maatregelen te nemen om te zorgen voor verplichte openbaarmaking door vennootschappen van onder meer de oprichtingsakte,

de statuten, de benoeming en de beëindiging van hun ambt. De verordening betreffende het vrije verkeer van niet-persoonsgegevens, om een laatste voorbeeld te geven, ontmoedigt zowel organisaties in de publieke sector als de particuliere sector om niet-persoonsgegevens te privatiseren.

Op het gebied van open data hebben de afgelopen jaren drie belangrijke ontwikkelingen plaatsgevonden:

1. Digitalisering: overheidsdocumenten lagen vroeger in archieven, bibliotheken of speciaal daarvoor bestemde documentatiecentra. Tegenwoordig worden steeds meer documenten online beschikbaar gesteld. Dit heeft een belangrijk effect op de zogenaamde 'practical obscurity'. Het feit dat men zich in het verleden de moeite moest getroosten om naar de plaats te gaan waar de documenten waren opgeslagen, ze op te vragen en in te zien, betekende dat in de praktijk slechts een beperkt aantal mensen de informatie zou raadplegen. In grote lijnen waren dat journalisten, historici, kritische burgers die de overheid op de voet volgden en amateurhistorici die hun stamboom onderzochten. Door de documenten openbaar te maken op het internet en geen toegangsbarrières op te werpen kan iedereen deze documenten gemakkelijk bekijken.
2. Actieve openbaarmaking: in het pre-digitale tijdperk werden de meeste documenten 'passief openbaar gemaakt'; burgers, journalisten en anderen kregen op verzoek toegang tot bepaalde documenten. Ze moesten dus al een globaal idee hebben van wat ze zochten, de openbaarmaking van documenten vereiste hun initiatief en de documenten werden meestal slechts voor een bepaalde periode beschikbaar gesteld. Momenteel worden documenten steeds vaker actief openbaar gemaakt; de overheid publiceert documenten niet op verzoek, maar op eigen initiatief. Dit betekent dat er geen specifieke reden meer is waarom een document beschikbaar wordt gesteld. Iedereen heeft er toegang toe en op elk moment.
3. Technologieën: de technische mogelijkheden om dergelijke documenten te doorzoeken zijn aanzienlijk toegenomen. Deze omvatten algoritmen en kunstmatige intelligentie die teksten kunnen analyseren op woorden, correlaties en onderwerpen. Waar het voorheen vooral individuen waren die toegang zochten tot overheidsdocumenten, zijn het momenteel technologiebedrijven die de beste uitgangspositie hebben om de miljoenen overheidsdocumenten die online verschijnen te scannen en te analyseren.

Daarnaast heeft er, gezien de algemene beschikbaarheid van data en datatechnologieën, het gemak van dataverzameling en -verwerking en de lagere kosten, een belangrijke verschuiving plaatsgevonden in het type dataverwerking. Gezien de kosten en praktische en technologische beperkingen voor het verzamelen van gegevens, waren veel gegevensoperaties, zelfs tot 20 jaar geleden, heel doelgericht. Er was een specifiek en vooraf vastgesteld doel waarvoor specifieke gegevens over specifieke entiteiten werden verzameld. Momenteel hebben echter veel, zo niet de meeste, gegevensverwerkingsoperaties betrekking op structurele en systemische gegevensverzamelingen, zoals camera's en sensoren die iedereen, overal in het publieke domein permanent bewaken en alomtegenwoordige online tracking. Deze verschuiving betekent dat de verzamelde gegevens vaak niet betrekking hebben op vooraf geïdentificeerde individuen, maar op groepen, categorieën of de gehele bevolking. Dit heeft op zijn beurt een verschuiving in gang gezet van de analyse van individuele gegevens naar die van statistische en geaggregeerde gegevens, van directe naar afgeleide gegevens en van zekere naar probabilistische informatie.

Deze ontwikkelingen hebben effect op de manier waarop de huidige wetgeving inzake gegevensbescherming is ingericht en met name op de categorale aanpak.

1. Werken met afgebakende definities van verschillende soorten gegevens gaat alleen als een 'datum' op een relatief stabiele manier in één categorie valt. Dit is steeds minder het geval. De aard van de data in ‘Big Data’-processen is niet stabiel, maar veranderend. Een dataset met gewone persoonsgegevens kan worden gekoppeld aan, en verrijkt met, een andere dataset om gevoelige gegevens af te leiden; de gegevens kunnen vervolgens worden geaggregeerd of ontdaan van identificatiegegevens; vervolgens kunnen de gegevens worden gedeanonimiseerd

of geïntegreerd in een andere dataset om persoonsgegevens te creëren. Dit alles kan in een fractie van een seconde gebeuren. De vraag is dus of het zinvol is om met goed gedefinieerde categorieën te werken als dezelfde 'datum' of dataset letterlijk van seconde tot seconde in een andere categorie kan vallen.

2. Ook wordt het steeds moeilijker om de status van gegevens precies te bepalen. De beoordeling of de gegevens de identificatie van een persoon mogelijk maken en of de informatie al dan niet als anoniem kan worden beschouwd, hangt af van de omstandigheden van het geval. Om de huidige status van een datum of dataset te bepalen, moet daarom rekening worden gehouden met de verwachte toekomstige status van de gegevens. Gezien de algemene beschikbaarheid van technologieën en de minimale investering die nodig is, wordt het steeds waarschijnlijker dat wanneer een database wordt gedeeld of anderszins beschikbaar wordt gesteld, er een partij is die deze gaat verrijken met andere gegevens. Zo wordt het steeds waarschijnlijker dat als een geanonimiseerde dataset openbaar wordt gemaakt, er een partij is die deze de-anonimiseert of combineert met andere data om persoonlijke profielen te maken; dat als een set persoonsgegevens wordt gedeeld, er een partij is die de gegevens zodanig gaat gebruiken om te komen tot een dataset met gevoelige persoonsgegevens; enzovoort. Aan de andere kant zullen er andere partijen zijn die toegang hebben tot die gegevens, maar zich niet bezighouden met dergelijke activiteiten; partijen die de gegevens niet zullen gebruiken, gebruiken zoals deze worden verstrekt of zelfs een database met persoonsgegevens de-identificeren. Wie wat doet is vooraf niet duidelijk. De juridische categorie waartoe de gegevens behoren is dus niet langer een kwaliteit van de gegevens zelf, maar een product van de inspanningen en investeringen van een verwerkingsverantwoordelijke.
3. De vraag is of het onderscheid tussen verschillende categorieën gegevens nog relevant is. De achterliggende gedachte is dat de verwerking van persoonsgegevens gevolgen heeft voor natuurlijke personen, terwijl de verwerking van niet-persoonsgegevens dat niet heeft en dat de verwerking van gevoelige persoonsgegevens zeer grote gevolgen kan hebben (groter dan de verwerking van ‘gewone’ persoonsgegevens gegevens normaal gesproken heeft), zodat dit onder het strengste regime valt, persoonsgegevens onder het 'normale' beschermingsregime vallen en de verwerking van niet-persoonsgegevens aan geen enkele beperking onderworpen is. De vraag is in hoeverre deze aanname nog houdbaar is in de 21e eeuw. Moderne gegevensverwerking op basis van geaggregeerde gegevens kan grote individuele en maatschappelijke gevolgen hebben. Profilering van groepen in plaats van individuen betekent voorts dat de gevolgen aanzienlijk kunnen zijn, maar niet altijd direct te relateren zijn aan individuen.

4. De impact van huidige en toekomstige datatechnologieën op de juridische categorieën

Dit onderzoek richtte zich op de technologische ontwikkelingen met betrekking tot vier toepassingsgebieden, namelijk anonimisering, aggregatie, pseudonimisering en het afleiden van gevoelige gegevens uit niet-gevoelige (persoons)gegevens. De bevindingen met betrekking tot elk van deze toepassingsgebieden zullen hieronder worden samengevat.

Anonimisering

De meest relevante anonimiseringstechnieken in het kader van dit onderzoek zijn:

1. Maskeren: beoogt een relatie te genereren tussen de oorspronkelijke set X en de gegenereerde set Y, zodat de indirecte identifiers worden gemaskeerd.
   1.1 Niet-perturbatieve maskering: gedeeltelijke onderdrukking of reductie van detail of verruwing van de originele dataset X. Hierdoor is dataset Y niet per se een verstoorde dataset, maar eerder een gereduceerde versie van de dataset X. Niet-perturbatieve maskering omvat onder andere:
      1.1.1 Sampling: vrijgeven van een sample S van de originele dataset X. Sampling is

geschikt voor kwalitatieve identifiers waarop geen rekenkundige bewerkingen kunnen worden uitgevoerd, zoals de oogkleur van een persoon of de maanden van het jaar;

1.1.2 Generalisatie: reductie van data granulariteit zodat dataset Y minder nauwkeurig is dan dataset X. Deze techniek is geschikt voor kwalitatieve identifiers, omdat het de maskering van bestanden met ongebruikelijke combinaties ondersteunt;

1.1.3 'Top- en bottom'-codering: een speciaal geval van generalisatie waarbij top-codes of bottom-codes worden ingesteld vanuit de originele identifiers van dataset X;

1.1.4 Onderdrukking: verwijdering van de gehele of van bepaalde identifiers in dataset Y vóór de vrijgave ervan. Aangezien het herstellen van informatie niet mogelijk is, wordt onderdrukking beschouwd als de sterkste anonimiseringstechniek.

1.2 Perturbatieve maskering: de vervorming of verstoring van microdata zodat de statistische eigenschappen van de oorspronkelijke dataset X behouden blijven in dataset Y. Perturbatieve maskering omvat onder meer:

1.2.1 Ruistoevoeging: maskering van identifiers door willekeurige ruis toe te voegen;

1.2.2 Data swapping: het uitwisselen van identifiers tussen individuele records;

1.2.3 Microaggregatie: clustering van records van dataset X in kleine aggregaten of groepen van k elementen, waarbij het gemiddelde van de waarden van de groep waartoe het record behoort, wordt gepubliceerd in dataset Y.

2. Synthetische data: heeft tot doel een dataset Y te creëren die bestaat uit willekeurig gesimuleerde bestanden die niet direct uit de dataset X zijn afgeleid, met behoud van de statistische eigenschappen van de oorspronkelijke dataset X. Als zodanig kunnen standaarddeviaties, medianen, lineaire regressie of andere statistische technieken worden gebruikt om synthetische gegevens te genereren.

Manieren om anonimiteit vanuit een technisch perspectief te definiëren omvatten, maar zijn niet beperkt tot:

1. k-anonimiteit: probeert de her-identificatie van bestanden te voorkomen op basis van een vooraf gedefinieerde set van indirecte identifiers. Een cel in een database verwijst in ieder geval naar k individuen;
2. l-diversiteit: heeft tot doel ervoor te zorgen dat elke groep gevoelige identifiers verschillende waarden bevat en dat geen van deze waarden domineert in frequentie;
3. t-closeness: stelt het gebruik van een relatief instrument voor om de variabiliteit van de waarden van de gevoelige identifiers te meten, waardoor de informatiewinst over de betrokkenen wordt beperkt. Alle waarden die door het sensitieve attribuut worden aangenomen, worden als even gevoelig beschouwd;
4. ε-differentiële privacy: de gegevensbeheerder genereert geanonimiseerde weergaven van een dataset met behoud van een kopie van de originele gegevens. Die weergaven of subsets zijn dus anoniem, maar de gegevensbeheerder heeft vaak nog steeds identificerende informatie.

Hoewel elk van deze technieken waardevol is, kan geen van deze technieken absolute anonimiteit garanderen. Met voldoende tijd, middelen en adequate technologie kunnen vrijwel alle geanonimiseerde gegevens worden ge-deanonimiseerd. Al in 2009 concludeerde Paul Ohm dat data ofwel waardevol ofwel perfect anoniem kunnen zijn, maar nooit beide. Technische literatuur onderstreept dat dit punt nu meer dan ooit waar is. Technische experts verwachten, zoals blijkt uit de interviews, geen revolutionaire nieuwe ontwikkelingen op het gebied van anonimisering of de-anonimisering, maar menen over het algemeen dat volledige anonimisering, zeker in juridische zin, steeds moeilijker zal worden gezien de algemene beschikbaarheid van technologieën en de algemene beschikbaarheid van gegevens.

Aggregatie

Door aggregatie worden de gegevens in een dataset niet op individueel niveau (n = 1) gepresenteerd, maar op geaggregeerd niveau (n = 10; n = 100; n = 1000). Hoe hoger het aggregatieniveau, hoe waarschijnlijker het is dat de dataset juridisch gezien geen persoonsgegevens bevat, hoewel een dergelijke beoordeling altijd afhankelijk is van de omstandigheden van het geval. De meest relevante aggregatietechnieken in het kader van dit onderzoek zijn:

1. Aggregatie op basis van derden: vertrouwde derden kunnen onbewerkte gegevens verzamelen, deze gegevens aggregeren en de resulterende gegevens overdragen aan geautoriseerde ontvangers. Op deze manier hebben de ontvangers alleen de geaggregeerde gegevens. Dit is echter mogelijk niet het geval voor een vertrouwde derde partij.
2. Aggregatie op basis van gegevensverstoring: willekeurige ruis wordt toegevoegd aan de verzamelde gegevens zodat de oorspronkelijke gegevens niet traceerbaar zijn, maar geaggregeerde waarden kunnen nog steeds worden berekend met een kleine of verwaarloosbare fout. Het nadeel van gegevensverstoring is het verschil tussen de oorspronkelijke gegevens en de verstoorde gegevens, wat in bepaalde gevallen kan leiden tot ongelijkheden in de berekening.
3. Aggregatie op basis van cryptografie: cryptografische primitieven kunnen worden gebruikt om de nadelen van de vorige methoden weg te nemen. Volledig homomorfe encryptie is een encryptietechnologie waarmee analyses in de cijfertekst op dezelfde manier als in de leesbare tekst kunnen worden uitgevoerd zonder de geheime sleutel te delen. Dit houdt in dat de berekening wordt uitgevoerd over de versleutelde gegevens zonder de noodzaak om deze te ontsleutelen, waardoor het delen van gegevens met derden mogelijk wordt. De resultaten van de berekening zijn gelijkelijk versleuteld, zodat alleen de data-exporteurs de gegevens kunnen ontsleutelen.
4. Statistical Disclosure Control: misschien wel de belangrijkste techniek voor het verzamelen van gegevens, vooral in het licht van het openbaar maken van de gegevens, is Statistical Disclosure Control (SDC). SDC heeft als doel om, zowel direct als indirect, identificerende informatie in een dataset te elimineren, met zoveel mogelijk behoud van de datakwaliteit. De specialist die verantwoordelijk is voor het beschermen van de gegevens moet verschillende methoden van openbaarmakingscontrole gebruiken, zodanig dat het minimaal vereiste beschermingsniveau wordt bereikt en dat het informatieverlies zo klein mogelijk is, wat per situatie zal verschillen. Wat informatieverlies is, kan niet als zodanig worden bepaald, omdat informatie een subjectief begrip is dat door elke gebruiker anders kan worden gedefinieerd.

Hoewel de technische mogelijkheden voor het anonimiseren van gegevens in geaggregeerde datasets groot zijn, en in het algemeen groter dan wanneer gegevens niet worden geaggregeerd, doet zich een nieuw probleem voor, dat in de technische literatuur wordt aangeduid als het samenstellingsprobleem. Dit betekent dat uit de combinatie van twee of meer datasets die zelf geen persoonsgegevens bevatten, persoonsgegevens kunnen worden afgeleid. Het kan gaan om gegevens over geïdentificeerde personen die vroeger in die databases zaten, maar het kan ook om andere personen gaan. Bovendien moet worden benadrukt dat als een partij algemene informatie zou gebruiken om beslissingen te nemen die van invloed zijn op personen, dit op juridisch gebied als persoonsgegevens zou worden gekwalificeerd. Uiteraard is het vooraf moeilijk in te schatten welke partij welke geaggregeerde data zal gebruiken voor welk type besluitvorming.

Hoewel anonimisering van geaggregeerde gegevens in isolatie potentieel mogelijk is, als bijvoorbeeld alleen de dataset als een relevante bron voor identificatiedoeleinden wordt beschouwd, zijn zowel de literatuur als de voor dit onderzoek geïnterviewde experts het erover eens dat dit steeds minder bepalend zal zijn. Dat heeft niet zozeer te maken met ontwikkelende technieken, maar met het groeiende datalandschap en de beschikbaarheid van open data. Omdat het waarschijnlijk is dat bijna elke geaggregeerde dataset op termijn zal worden gebruikt voor gevolgtrekkingen op persoonlijk niveau, voor samenstellingsactiviteiten en/of voor het ontwikkelen van besluitvormingsbeleid dat gevolgen heeft voor mensen, kan, vanuit juridisch perspectief, geen enkele geaggregeerde dataset worden aangemerkt als absoluut buiten het gegevensbeschermingsregime vallend.

Pseudonimisering

De meest relevante pseudonimiseringstechnieken in het kader van dit onderzoek zijn:

1. Hashing is een techniek waarmee pseudoniemen kunnen worden afgeleid. In een notendop zijn hashfuncties functies die een invoer van willekeurige lengte comprimeren tot een resultaat met een vaste lengte. Deze uitvoer met een vaste grootte wordt een berichtsamenvatting, hash-waarde, hash-code of gewoon hash genoemd. Als een identifier *m* wordt gebruikt als invoer in de hash-functie *h*, zal de functie een pseudoniem met vaste grootte *h(m)* teruggeven.
2. Hashing met een sleutel of keyed hashing bouwt voort op conventionele hashing door een geheime sleutel toe te voegen die de uitvoer van de functie *h* verandert. Hashing met sleutel kan verschillende pseudoniemen produceren voor dezelfde invoer, afhankelijk van de keuze van de specifieke sleutel.
3. 'Salted' hashing is een variant van keyed hashing, waarbij gebruik wordt gemaakt van een conventionele hashfunctie in combinatie met een zogenaamde 'salt', of aanvullende willekeurig uitziende data. Net als keyed hashing, produceert hashing met salt verschillende pseudoniemen voor dezelfde initiële identifier. Daarom heeft salted hashing dezelfde eigenschappen als keyed hashing, zolang het 'salt' op de juiste manier is beveiligd en derden er geen kennis van hebben.
4. Peppered hashing bestaat uit het toevoegen van een geheim aan het 'salt' tijdens het hashen en het apart opslaan van salts en pseudoniemen in een ander medium, bijvoorbeeld in een hardware beveiligingsmodule. De 'pepper' deelt daarom bepaalde eigenschappen met salt omdat het een willekeurige waarde is en vergelijkbaar met een coderingssleutel omdat het geheim moet worden gehouden.
5. Tokenisatie bestaat uit het vervangen van identifiers door willekeurig gegenereerde waarden, ook wel tokens genoemd, zonder enige wiskundige relatie en zonder het type of de lengte van de gegevens te veranderen. Dit is een belangrijk verschil met encryptie. In tegenstelling tot de laatstgenoemde, voorkomt de onveranderlijkheid van gegevenstypen en lengtes bij tokenisatie elke onbegrijpelijkheid van informatie door verwerking in tussenliggende systemen. Tegelijkertijd betekent dit ook een afname van de rekenkracht die nodig is om de tokens te verwerken. Aangezien er geen sleutels of algoritmen zijn gebruikt om de oorspronkelijke identifier uit het token af te leiden, impliceert de kennis van een token niet de openbaarmaking van persoonsgegevens.

De meest relevante encryptietechnieken in het kader van dit onderzoek zijn:

1. Symmetrische encryptie bestaat uit het gebruik van één geheime sleutel om elektronische informatie zowel te versleutelen als te ontsleutelen. Partijen die vertrouwen op symmetrische versleuteling moeten de geheime sleutel delen om het ontsleutelingsproces mogelijk te maken. Symmetrische encryptie transformeert de initiële identifier (maar ook de volledige dataset) in een pseudoniem (of cijfertekst), die vervolgens wordt gedecodeerd om de initiële identifier te onthullen.
2. Asymmetrische encryptie bestaat uit het gebruik van twee sleutels, een openbare en een privésleutel, om elektronische informatie zowel te versleutelen als te ontsleutelen. Partijen die vertrouwen op asymmetrische versleuteling moeten vertrouwen op de openbare sleutel om de gegevens te versleutelen en op de privésleutel om deze te ontsleutelen. Openbare en privésleutels zijn wiskundig gerelateerd, maar worden op passende wijze onderscheiden door de introductie van willekeur in het coderingsproces om te voorkomen dat de privésleutel kan worden vastgesteld.
3. Homomorfe encryptie maakt berekeningen op versleutelde gegevens mogelijk. Berekenen op versleutelde gegevens verwijst naar het feit dat een partij *Pn* die de initiële identifiers of invoer *mn* heeft en de functie *f* wil berekenen om $f(m1,\ldots,mn)$ te verkrijgen, in plaats daarvan de versleuteling of pseudoniemen van de invoer *cn* kan berekenen om $f'(c1,\ldots,cn)$ te verkrijgen, die ontcijferd kan worden tot $f(m1,\ldots,mn)$. Het voordeel van homomorfe encryptie is dat

persoonsgegevens vertrouwelijk blijven terwijl ze worden geanalyseerd of 'gemined' zonder dat ze moeten worden ontsleuteld en de uitvoer in gevaar komt.

4. Multiparty computation (MPC) verschilt van de drie eerder besproken technieken, hoewel het gerelateerd is aan homomorfe encryptie. MPC is een techniek die zich toelegt op protocollen waarmee een reeks partijen gezamenlijk een functie van hun invoer of identificatiegegevens kan berekenen, terwijl wordt vermeden dat iets anders wordt onthuld dan de uitvoer van de genoemde functie. MPC zorgt ervoor dat de input van de partijen in het algemeen geheim blijft tijdens de gehele verwerking van data-aggregatie, en wordt dus beschouwd als een geavanceerd privacy-behoudend instrument voor pseudonimisering. Het kan worden gebruikt als een encryptietechniek, maar is veel breder in termen van mogelijke toepassingen.

In de technische literatuur worden encryptie- en pseudonimiseringstechnieken meestal beschreven als vormen van privacyverhogende of -behoudende technologieën. Welke techniek het meest geschikt is, hangt af van de context, het type gegevens, de betrokken actoren en andere aanwezige waarborgen. Dat is de reden waarom, hoewel sommige technieken over het algemeen als zwakker worden beschouwd dan andere, van geen enkele techniek kan worden gezegd dat deze de voorkeur heeft, en geen enkele techniek categorisch kan worden uitgesloten. Sommige pseudonimiserings- of encryptietechnieken, vooral wanneer ze worden toegepast in combinatie met andere privacyverhogende technologieën, kunnen zo sterk zijn dat ze de belangen van betrokkenen beter kunnen beschermen dan bepaalde anonimiseringstechnieken.

Inferentie van gevoelige gegevens

Uit de technische literatuur blijkt duidelijk dat het steeds gemakkelijker wordt om persoonsgegevens af te leiden uit geaggregeerde gegevens en gevoelige persoonsgegevens uit persoonsgegevens of niet-persoonsgegevens. Statistische organisaties en volkstellingsbureaus publiceren bijvoorbeeld vaak geaggregeerde datasets, die volgens hen geen persoonlijke informatie bevatten. Maar door middel van zogenaamde databasereconstructie-aanvallen is het vaak mogelijk om het geslacht, de leeftijd, het ras, de etniciteit en gedetailleerde geografische locatie te reconstrueren die is vastgelegd voor ongeveer de helft van de bevolking in de dataset. Bovendien kunnen door het combineren van twee datasets die zelf geen persoonsgegevens bevatten, persoonsgegevens worden verkregen en zelfs gevoelige persoonsgegevens worden afgeleid. Bijgevolg kunnen zowel door de beschikbaarheid van open data als door de toegenomen technologische capaciteiten om data af te leiden, gevoelige gegevens worden gedestilleerd uit zowel 'gewone' persoonsgegevens als uit niet-persoonsgegevens. Zowel wetenschappelijke literatuur als de voor dit onderzoek geïnterviewde experts benadrukken dat deze trend in de loop der tijd alleen maar zal toenemen.

5. De kloof tussen het wettelijke regime en de technologische realiteit

Er zijn verschillende spanningsvelden tussen het technologische domein en de manier waarop het wettelijk kader is opgesteld. De in het kader van dit onderzoek belangrijksten zullen hieronder worden toegelicht.

Anonieme gegevens

1. Hoewel het wettelijk kader onderscheid maakt tussen anonieme en pseudonieme gegevens, is dit onderscheid voor technische experts niet onomstreden. Vanuit technisch oogpunt zouden data anoniem genoemd kunnen worden wanneer een aantal relevante variabelen worden verwijderd. Veel technische experts gaan uit van niveaus van anonimiteit. Er is een schaal van volledige anonimiteit naar directe identificeerbaarheid in plaats van een binair onderscheid, zoals is vervat in de AVG.
2. Het feit dat de AVG geen tijdslimiet stelt aan wanneer gegevens opnieuw kunnen worden

geïdentificeerd of geanonimiseerd, betekent dat het zeer waarschijnlijk is dat de gegevens op een bepaald moment aan een natuurlijke persoon zullen worden gekoppeld en dus als persoonsgegevens dienen te worden aangemerkt.
3. Een aantal technische experts vraagt zich af of de wettelijke definitie van anonieme gegevens in de 21e eeuw kan worden gehandhaafd, aangezien het steeds moeilijker zal worden om aan de wettelijke drempel te voldoen. Vanuit technologisch oogpunt is het bijna onmogelijk om over echt anonieme gegevens te beschikken. Met name wanneer geanonimiseerde datasets worden gedeeld of online beschikbaar worden gesteld, is de kans groot dat er een partij is die de data her-identificeert of samenvoegt met andere datasets om tot persoonsgegevens te komen.
4. Sommige auteurs concluderen dat de stand van de techniek die verband houdt met de technieken die door de Artikel 29 Werkgroep zijn opgesomd, bevestigt dat anonimiseringsmethoden voor grote uitdagingen staan met betrekking tot de originele gegevens en dat dit niet langer vanuit een statisch perspectief kan worden beschouwd, maar een dynamisch perspectief vereist.
5. Veel technische experts vinden de juridische definitie van anonimisering onduidelijk en vaag. Een bijzonder punt van aandacht is de term 'redelijkerwijs valt te verwachten'.

Geaggregeerde gegevens

1. Wettelijke regelgeving behandelt (micro)data en geaggregeerde (macro)data hetzelfde, terwijl er wel duidelijk verschillende risico's verbonden zijn aan het openbaar maken van micro- en macrodata. Op datasetniveau vereist het spreken van absoluut anonieme gegevens in het algemeen dat de dataset zodanig wordt gestript dat er vrijwel geen relevante informatie overblijft, terwijl er op geaggregeerd niveau veel meer mogelijkheden zijn om individuen te beschermen tegen identificatie. Op hun beurt kunnen geaggregeerde gegevens op andere manieren aan personen worden gekoppeld, met name wanneer deze online beschikbaar worden gesteld.
2. De toegenomen beschikbaarheid van open data maakt het moeilijk om een goede inschatting te maken van de risico's die gepaard gaan met het vrijgeven van geaggregeerde gegevens door statistische organisaties of andere partijen.
3. Statistische gegevens worden gebruikt om kennis te genereren door middel van analyse van bestaande data om zo aannames over individuen te doen, bijvoorbeeld, door het in kaart brengen van ervaringen uit het verleden en het vastleggen van correlaties tussen bepaalde karakteristieken, bepaalde uitkomsten, of bepaald gedrag. Met AI- en Big Data-analyse kunnen mensen op bruikbare manieren worden geprofileerd zonder persoonlijk of individueel te worden geïdentificeerd. Aangezien de stand van de technologie het mogelijk maakt om meer informatie uit niet-persoonsgegevens te halen, wordt een grotere rol toegekend aan het gebruik van dergelijke gegevens. Deze trend kan niet adequaat worden aangepakt door het gegevensbeschermingskader, dat sterk is gebaseerd op de notie van de identificatie van individuele natuurlijke personen.
4. Voor veel technische experts en professionals geeft het juridische regime tegenstrijdige signalen. Enerzijds worden open data, hergebruik van overheidsinformatie en dataportabiliteit bevorderd; anderzijds wordt de nadruk gelegd op privacy, geheimhouding en gegevensbescherming. Wat deze spanning complexer maakt, is dat wetgevers en rechters geen uniform beeld geven van in hoeverre het verzamelen en gebruiken van geaggregeerde gegevens gereguleerd moet worden of in hoeverre geaggregeerde gegevens ook persoonsgegevens kunnen zijn.

Pseudonieme gegevens

1. Technische experts gaan er traditioneel van uit dat pseudonimisering betekent dat een of meer identifiers worden vervangen door een pseudoniem. De AVG definieert pseudonimisering echter als verwerking waarbij aanvullende informatie, die her-identificatie mogelijk maakt, op een andere plaats wordt opgeslagen; er kunnen pseudonieme gegevens zijn zonder een expliciet

pseudoniem. De twee definities liggen zeer dicht bij elkaar, maar zijn niet volledig identiek.
2. Technisch gezien is het niet duidelijk waarom pseudonimisering, als vorm van risicopreventie, een bijzondere status zou moeten hebben binnen het wettelijk kader, omdat er meerdere manieren en technieken zijn om dit te bewerkstelligen. De voorkeur geven aan deze ene techniek lijkt in schril contrast te staan met de veronderstelde technologische neutraliteit van het gegevensbeschermingskader.
3. Niet alle vormen van pseudonimisering zijn even veilig. Het wettelijke regime geeft geen richtlijnen over welk type techniek het meest geschikt is voor welk type context.

Gevoelige persoonsgegevens

1. Experts zetten vraagtekens bij de vaste categorieën gevoelige gegevens die in de AVG worden gebruikt. Zo zouden de financiële positie, de sociaaleconomische achtergrond of inkomen in veel gevallen als gevoelige gegevens kunnen worden behandeld, omdat de potentieel nadelige effecten van het verwerken van dergelijke gegevens op zijn minst even ernstig kunnen zijn als het verwerken van, bijvoorbeeld, het lidmaatschap van een politieke organisatie of een vakbond. Zij wijzen erop dat wat als gevoelige persoonsgegevens moet of kan worden beschouwd, verschilt per regio of land. Daarom kan het werken met één vaste lijst van soorten gevoelige gegevens voor alle EU-landen bijzonder uitdagend zijn.
2. Het wettelijk regime richt zich op vaste categorieën gegevens, terwijl wat technisch wel of niet gevoelig is, niet afhankelijk is van het type gegevens. Gegevensverwerking kan gevoelig en schadelijk zijn, zelfs zonder de verwerking van de categorieën gegevens die in de AVG worden vermeld, of kan niet-schadelijk zijn, zelfs als een of meer van de soorten gegevens die als gevoelig zijn gecategoriseerd, worden verwerkt.
3. Veel technische experts wijzen op het feit dat gevoelige informatie vaak kan worden afgeleid uit niet-gevoelige persoonsgegevens en zelfs uit niet-persoonsgegevens. Het in het wettelijke regime gehanteerde binaire onderscheid tussen gevoelige en niet-gevoelige gegevens houdt onvoldoende rekening met de technologische complexiteit en realiteit op dit punt.

6. Reguleringsalternatieven gevonden in wet en literatuur

Er zijn alternatieven voorgesteld voor het huidige regulerende regime. De meest relevante voor de doeleinden van dit onderzoek zijn:

Anonieme gegevens

1. Maak een einde aan het onderscheid tussen anonieme en niet-anonieme gegevens. Als het steeds waarschijnlijker wordt dat gegevens worden gedeanonimiseerd en als niet-persoonsgegevens kunnen worden gebruikt voor ingrijpende gegevensprocessen, kan de keuze om anonieme of niet-persoonsgegevens buiten de reikwijdte van de gegevensbeschermingswetgeving te plaatsen overbodig worden.
2. Gebruik een minder breed concept dan de huidige definitie van persoonsgegevens om een duidelijker onderscheid te maken tussen persoonsgegevens en geanonimiseerde gegevens. Het begrip ‘identificeerbaar’ zou bijvoorbeeld kunnen worden geschrapt, of er kan een specifieke horizon of tijdslimiet aan worden toegevoegd.
3. Creëer verschillende niveaus van identificeerbaarheid, wat betekent dat de toepassing van het gegevensbeschermingskader geen zwart-witkwestie is, maar een geleidelijke schaal, bijvoorbeeld des te meer gegevens worden geanonimiseerd des te minder normen voor gegevensbescherming van toepassing zijn.
4. In plaats van te werken met het zeer contextuele 'alle middelen die redelijkerwijs door de verwerkingsverantwoordelijke of door een andere persoon kunnen worden gebruikt', zou kunnen worden overwogen om de AVG aan te passen met een zin die werd voorgesteld in het

wetgevingsproces van de Richtlijn bescherming persoonsgegevens, namelijk 'ten koste van een buitensporige inspanning'.

Geaggregeerde gegevens

1. In plaats van een compromis te zoeken tussen open data en gegevensbescherming, zou een optie kunnen zijn om de gegevensbeschermingsregeling te laten prevaleren of het kader te laten bieden voor het gebruiken, delen en openbaar maken van statistische en geaggregeerde gegevens. Dit is in wezen wat het HvJ EU heeft gedaan in *Latvijas Republikas Saeima*.
2. Een nog verdergaande optie zou kunnen zijn om terug te keren naar een van de eerdere regels over statistische gegevens, namelijk dat ‘statistische gegevens alleen in geaggregeerde vorm mogen worden vrijgegeven als het onmogelijk is om de informatie aan een bepaalde persoon te koppelen.'
3. Er zou een uitgebreid kader kunnen komen om de behoefte aan open data en het verwerken van statistische gegevens enerzijds te verzoenen met de behoefte aan privacy en gegevensbescherming anderzijds. Een dergelijk kader zou op EU-niveau moeten worden aangenomen en niet aan de lidstaten moeten worden overgelaten, en zou in detail moeten specificeren hoe deze twee beginselen in concrete situaties met elkaar kunnen worden verenigd.
4. Het gegevensbeschermingskader zou explicieter kunnen zijn in termen van een drempel of grens van gegevensanonimiteit wanneer gegevens worden geaggregeerd of in termen van de technische normen die moeten worden toegepast bij het vrijgeven van geaggregeerde gegevenssets.
5. Een radicaler alternatief zou kunnen zijn om manieren te vinden om de regelgeving inzake privacy en gegevensbescherming te baseren op andere concepten dan identificeerbaarheid. Zo hebben sommige auteurs voorgesteld om de focus op individuele privacy en identificeerbaarheid van natuurlijke personen los te laten en in plaats daarvan of in aanvulling daarop meer nadruk te leggen op de bescherming van groepen, categorieën en datacollectieven.
6. Er kunnen concretere regels voor het openbaar maken van geaggregeerde gegevens worden ontwikkeld, zoals het hebben van een minimum *n* per cel met een frequentieteltabel of regels over dominantie met kwantitatieve magnitudetabellen. Ook zouden controles voor het vrijgeven van groepsdata kunnen worden gegeven.

Pseudonieme gegevens

1. De AVG of het Europees Comité voor gegevensbescherming (European Data Protection Board - EDPB) zou meer richtsnoeren kunnen geven ten aanzien van welke soorten pseudonimiseringstechnieken het meest geschikt worden geacht voor welke context.
2. Overwogen kan worden om pseudonimisering af te stemmen op het concept van gegevensbeheer. Een gegevensbeheerder kan fungeren als een pseudonimiseringsentiteit die verantwoordelijk is voor het verwerken van pseudoniemen, die onder specifieke voorwaarden gegevenstoegang kan verlenen aan onderzoekers of bedrijven en gegevens kan afschermen tegen ongewenste of onrechtmatige toegang.
3. Sommige deskundigen stellen voor om de specifieke verwijzing naar pseudonimisering uit de AVG te schrappen, zowel omdat het als te vaag wordt beschouwd als omdat er geen reden is om deze techniek te verkiezen boven andere risicomijdingstechnieken.
4. Anderen daarentegen hebben gesuggereerd om de categorie pseudonieme gegevens een nog prominentere rol te geven, waardoor het een officiële tussencategorie wordt tussen anonieme gegevens en persoonsgegevens.

Gevoelige persoonsgegevens

1. Verschillende auteurs hebben gesuggereerd dat de gevoeligheid van gegevensverwerking niet

langer afhangt van het type gegevens dat wordt verwerkt, maar veeleer van de verwerkingstechnologieën en het gebruik ervan. Daarom hebben zij voorgesteld om de bijzondere regeling voor gevoelige persoonsgegevens uit het gegevensbeschermingskader weg te laten.
2. Verruim de lijst met gevoelige gegevens en neem daarin o.a. financiële gegevens op, zoals bij het opstellen van de AVG werd gesuggereerd maar uiteindelijk werd afgewezen.
3. Als alternatief is voorgesteld om te werken met een lijst met voorbeelden in plaats van met vaste categorieën, wat de oorspronkelijke benadering was voor de regulering van gevoelige persoonsgegevens. Ook zou kunnen worden overwogen om een restcategorie in te voeren, vergelijkbaar met de verwijzing naar 'of andere status' in artikel 14 van het Europees Verdrag voor de Rechten van de Mens (EVRM).
4. Er zou kunnen worden overwogen om onderscheid te maken tussen de verschillende categorieën gevoelige persoonsgegevens, een benadering die lijkt te worden gevolgd door het HvJ EU, waarbij gezondheidsgegevens in de meest gevoelige categorie worden geplaatst, terwijl andere gegevens als minder gevoelig kunnen worden beschouwd.
5. Hoewel de meeste zorgen gaan over de vraag of de AVG strikt genoeg is voor speciale categorieën gegevens, zijn er ook tegengestelde argumenten. Er is een voortdurende discussie in hoeverre het mogelijk is om gevoelige persoonsgegevens te verwerken om discriminatie te voorkomen, bijvoorbeeld in KI-systemen. Het gebruik van gevoelige persoonsgegevens kan nodig zijn om discriminatie te voorkomen, vooral als het gaat om data gedreven besluitvorming. Om een van de onderliggende grondgedachten van de categorie bijzondere gegevens te bevorderen, namelijk het voorkomen van discriminerende praktijken, kan het dus nodig zijn om meer gevoelige persoonsgegevens te verwerken in plaats van minder.

7. Reguleringsdoelstelling van de gegevensbeschermingsregeling

Om te beoordelen of er lacunes in de regelgeving zijn en, zo ja, welke, is het nodig om te beoordelen wat de reguleringsdoelstelling van het privacy- en gegevensbeschermingsregime eigenlijk is. Dit is een punt van discussie: moet gegevensbescherming worden gezien als een regime dat grenzen en beperkingen oplegt aan verwerkingsverantwoordelijken of gaat het primair om de controlerechten van betrokkenen?

Enerzijds kan worden verwezen naar artikel 5 van de Algemene Verordening Gegevensbescherming, dat als de ruggengraat van deze wet wordt gezien. Het stelt dat persoonsgegevens rechtmatig, behoorlijk en op transparante wijze moeten worden verwerkt, moeten worden verzameld voor gespecificeerde, expliciete en legitieme doeleinden en niet verder moeten worden verwerkt op een manier die onverenigbaar is met die doeleinden, adequaat, relevant en beperkt moeten zijn tot wat nodig is met betrekking tot de doeleinden waarvoor ze worden verwerkt, nauwkeurig en, waar nodig, actueel moeten zijn, niet langer dan nodig worden bewaard in een vorm die identificatie van betrokkenen mogelijk maakt, en verwerkt op een wijze die zorgt voor een passende beveiliging van de persoonsgegevens. Dit zijn allemaal verplichtingen die op de verwerkingsverantwoordelijke rusten en die van toepassing zijn onafhankelijk van eventuele rechten die door betrokkenen worden ingeroepen. Aan de andere kant worden er in het gegevensbeschermingsregime steeds meer rechten toegekend aan betrokkenen. Bovendien is, vooral door Duitse invloed, het begrip informatieve zelfbeschikking steeds populairder geworden. Daarom stellen sommigen dat, in plaats van verplichtingen die aan de verwerkingsverantwoordelijken worden opgelegd, de rechten van de betrokkenen de kern vormen van het gegevensbeschermingsregime.

Wat de beoordeling van deze kwestie bemoeilijkt, is het feit dat het gegevensbeschermingsregime niet alleen een beschermend doel heeft, maar dat artikel 1 AVG ook als doel erkent om de verwerking van persoonsgegevens in de EU te faciliteren. Een van de expliciete doelstellingen van het gegevensbeschermingskader van 1995 was het wegnemen van belemmeringen voor de doorgifte van

persoonsgegevens binnen de Unie door één gemeenschappelijk niveau van gegevensbescherming vast te stellen. Vóór de richtlijn had elk land eigen normen voor gegevensbescherming, wat het gebruik en de doorvoer van persoonsgegevens belemmerde. Het aannemen van één EU-breed kader voor gegevensbescherming loste dit probleem op. De regels in de AVG verbieden zelden specifieke gegevensverwerkingen. In de meeste gevallen bevatten ze procedurele waarborgen en beginselen die een correcte gegevensverwerking garanderen. Gegevensbescherming kan dus worden gezien als het bevorderen van gegevensverwerking door het bieden van een algemeen kader.

Tot slot bevat de AVG veel expliciete uitzonderingen voor specifieke verwerkingen. De belangrijkste in het kader van dit onderzoek zijn die met betrekking tot de vrijheid van meningsuiting, archivering, statistisch onderzoek, open overheid en het hergebruik van overheidsinformatie. De EU heeft de keuze gemaakt om verder te gaan dan het bevorderen van openheid en transparantie ten aanzien van overheidspraktijken; het heeft het hergebruik van overheidsinformatie gestimuleerd. Toch maakt de Open Data Richtlijn duidelijk dat deze geen invloed heeft op de AVG; wat dit in de praktijk betekent wordt opengelaten.

Binnen de EU bestaat onduidelijkheid over hoe om te gaan met de conflicten tussen de verschillende regimes en de beschermingsdoeleinden die eraan ten grondslag liggen. Over het algemeen neemt de EU-regelgever instrumenten aan die zijn gebaseerd op duidelijke en afgebakende gegevenscategorieën, terwijl er in de rechtspraak wordt gekozen voor een contextuele benadering. Adviesorganen zoals de Artikel 29-Werkgroep en de EDPB propageren ook een flexibele benadering en hebben in de loop der tijd de reikwijdte van onder meer persoonsgegevens verruimd. Gerechtshoven hebben duidelijke grenzen gesteld wanneer regelgevers onderscheid tussen typen data gebruiken om lagere beschermingsniveaus in te voeren, zoals toen het HvJ EU de EU dataretentie-richtlijn nietig verklaarde. Een vergelijkbare benadering is waar te nemen met betrekking tot de verschuiving naar open data en het hergebruik van overheidsinformatie. Hoewel dit sterk wordt gestimuleerd door de EU-wetgever, zijn rechters terughoudender. Zo vroeg het HvJ zich af of het voor het beschermen of verbeteren van de verkeersveiligheid noodzakelijk was om toegang te verlenen tot gegevens over verkeersovertredingen. Het stelde vast dat het regime derden toegang gaf tot de informatie, zelfs als die derden andere doeleinden hadden dan die welke verband hielden met het vergroten van de verkeersveiligheid, wat niet was toegestaan.

8. Gevaren van over- en onderregulering

De moeilijkheid bij het beoordelen van het bestaan van lacunes in de regelgeving en de wenselijkheid van reguleringsalternatieven is dat eerst de discussie over de reguleringsdoel(en) van het privacy- en gegevensbeschermingskader moet worden beslecht, terwijl dat een punt van discussie blijft. Bovendien is er geen duidelijke voorkeur in reguleringsbenadering: een categorale, een contextuele of een hybride. Elk heeft zijn eigen voor- en nadelen. Het is dus zowel een kwestie van smaak of er lacunes in de regelgeving zijn en zo ja, wat dat in de toekomst betekent voor de rechtsbescherming van gegevens in brede zin. Bovendien houdt de keuze tussen verschillende reguleringsopties een keuze in waar het reguleringsprerogatief wordt geplaatst. Hoe meer duidelijkheid er wordt verschaft in het wettelijke regime, hoe meer het prerogatief bij de wetgevende macht wordt gelegd. Hoe meer een contextuele benadering wordt gevolgd, hoe meer de rechterlijke en/of de uitvoerende macht de juiste interpretatie van de regels per context moet geven. Het eerste heeft het voordeel van democratische legitimiteit, het tweede van praktische toepasbaarheid. Het eerste heeft het voordeel dat het rechtszekerheid biedt door voor alle situaties één benadering in te voeren; het tweede heeft het voordeel dat het in staat is om granulariteit te bieden in regulering.

Om een voorbeeld te geven, misschien is de essentiële vraag die deze studie oproept wel of het begrip 'persoonsgegevens' en de subcriteria 'identificeerbaarheid' en ‘middelen redelijkerwijs valt te verwachten’ moeten worden behouden, of dat niet-persoonsgegevens moeten worden beschermd,

bijvoorbeeld onder een AVG-light regime, of dat er een geleidelijke schaal naar identificatie moet worden ingevoerd. Die vraag hangt af van wat de reguleringsgrondgedachte van het gegevensbeschermingskader wordt geacht te zijn. Als het de individuele belangen van natuurlijke personen beschermt, is er geen directe noodzaak om ook de verwerking van geaggregeerde of anonieme gegevens te regelen. Om mogelijke schade aan te pakken die voortvloeit uit beleid en acties op basis van groepsprofielen kan het aan de rechterlijke macht worden gelaten om het regulerende regime zo te interpreteren dat deze schade wordt gedekt, hetzij op basis van de AVG, hetzij op grond van artikel 8 EVRM. Als het doel van de regelgeving is om de datamacht van organisaties in de publieke en private sector in te perken, dan is het logisch om ook beperkingen en eisen te stellen aan de verwerking van niet-persoonsgegevens, en zou het geen probleem zijn om het gegevensbeschermingsregime uit te breiden om ook de verwerking van niet-persoonsgegevens te dekken en de materiële reikwijdte ervan los te koppelen van de identificeerbaarheid van een natuurlijke persoon. Beide keuzes roepen bovendien de vraag op van de specificiteit van de regelgeving. De toezichthouder handhaaft tot nu toe een strikt reguleringsonderscheid tussen niet-persoonsgegevens en persoonsgegevens, maar in de praktijk is dit onderscheid moeilijk te maken. Rechters hebben bijgevolg de definitie van persoonsgegevens uitgebreid tot gegevens die steeds meer perifeer zijn aan natuurlijke persoon, terwijl de verwerkingsverantwoordelijken om meer aanwijzingen vragen om dat onderscheid te maken. Het gevaar van het intact laten van de huidige aanpak is dat verantwoordelijke data-organisaties aan de veilige kant blijven, terwijl anderen de grenzen van de wet oprekken. Bovendien geldt dat hoe minder duidelijkheid in de regelgeving wordt gegeven, hoe moeilijker het zal zijn om de regels te handhaven, omdat elke gegevensverwerking mogelijk een eigen rechtmatigheidsbeoordeling vereist. Wanneer de keuze wordt gemaakt om het huidige regulerende regime intact te laten, is het dus nog steeds de vraag of er meer reguleringsrichtsnoeren moeten worden gegeven aan verwerkingsverantwoordelijken over het maken van onderscheid tussen gegevenscategorieën.

Bovendien, wanneer de keuze wordt gemaakt om niet-persoonsgegevens te onderwerpen aan wetgeving kunnen opnieuw twee verschillende benaderingen worden gevolgd: een categorale en een contextuele. Ofwel handhaaft het reguleringsregime een onderscheid tussen niet-persoonsgegevens en persoonsgegevens, maar hecht het een ander reguleringsregime aan niet-persoonsgegevens, ofwel heft het deze differentiatie en mogelijk andere gegevensonderscheiden op, en maakt het regime het type regels en de regeldruk voor de verwerkingsverantwoordelijken afhankelijk van de beoordeling van de betrokken risico's per geval (danwel gerelateerd aan individuele, groeps- en/of maatschappelijke belangen).

Voor de kwestie van overregulering is het van belang in hoeverre het stimuleren van gegevensverwerkingen op dezelfde voet wordt geplaatst als het beschermingsdoel van het gegevensbeschermingsregime en hoe het doel om open data-omgevingen en het hergebruik van overheidsinformatie te bevorderen wordt beoordeeld. Moet het bevorderingsdoel worden gezien als een even belangrijke grondgedachte als het beschermingsdoel, of kan deze grondgedachte alleen worden bevorderd binnen de grenzen die voortvloeien uit het beschermingsdoel? Als dat laatste het geval is, is overregulering geen wezenlijk risico, terwijl het voorkomen van onderregulering het hoofddoel is. Als beide doelen echter op dezelfde voet worden geplaatst, heeft het bevorderen van het ene doel bijna per definitie gevolgen voor het andere. Vervolgens rijst de vraag welk type regulering het meest effectief is. Hoewel een contextueel kader de meeste ruimte lijkt te laten voor data-innovatie, pleiten verwerkingsverantwoordelijken op het eerste gezicht vaak expliciet voor meer duidelijkheid en zekerheid op het gebied van regelgeving, omdat ze bang zijn voor terugslag en investeringen die niet lonend zijn.

Een soortgelijk punt moet worden opgemerkt met betrekking tot de beschermende grondgedachte. Experts hebben benadrukt dat de benadering van de AVG, waarbij het verwerken van gevoelige gegevens in principe verboden is, steeds meer het doel mist dat het beoogt, namelijk het beschermen van individuen tegen schade. Om discriminerende praktijken in KI-systemen te voorkomen, kan het

nodig zijn om gevoelige persoonsgegevens te verwerken. Anderen hebben benadrukt dat het zelfs in bredere zin nodig kan zijn om niet te focussen op gegevensminimalisatie maar op *gegevensminimumisatie*, op de eis dat een minimumniveau van gegevens wordt verzameld, geanalyseerd en opgeslagen. Het is dus zelfs een kwestie van debat of het beschermingsdoel uit de AVG het best gediend is door beperkingen op dataprocessen op te leggen.

9. Hoe zullen de huidige en toekomstige technische ontwikkelingen de komende periode van invloed zijn op de AVG en rechtsbescherming in brede zin?

Het is duidelijk dat de technologische ontwikkelingen en de algemene beschikbaarheid van data nu en in de toekomst tot gevolg hebben dat anonimisering steeds moeilijker wordt. De status van data wordt steeds volatieler en wordt steeds minder een kenmerk van data en datasets zelf en steeds meer een effect van de inspanningen van de verwerkingsverantwoordelijke. De juridische categorieën zullen steeds meer fluïde en minder stabiel worden en één database kan juridisch verschillend zijn per partij die er toegang toe heeft. Een database die op zichzelf alleen niet-persoonsgegevens bevat, kan worden omgezet in een database met persoonsgegevens door deze te combineren met een andere database, kan worden gebruikt om gevoelige persoonsgegevens te verkrijgen, om het volgende moment weer te worden geaggregeerd en geanonimiseerd. Gezien deze trends en gezien de begrippen 'identificeerbaarheid' en 'alle middelen die redelijkerwijs kunnen worden gebruikt', zullen steeds meer gegevens, zo niet alle, onder het gegevensbeschermingskader vallen en moeten ze mogelijk zelfs worden behandeld als (potentiële) gevoelige persoonsgegevens, waardoor het strengste van alle regimes zou gelden.

Of dit als problematisch wordt beschouwd, staat ter discussie en hangt af van wat de grondgedachte van het gegevensbeschermingskader is en welke effecten van onder- en overregulering het meest waarschijnlijk zijn. In dit onderzoek zijn geen verschillende scenario's gevonden voor hoe het technologische domein en de beschikbaarheid van open data zich in de loop van de tijd zullen ontwikkelen. Literatuur, geïnterviewde experts en experts die zijn uitgenodigd voor de workshop die voor dit onderzoek is gehouden, wijzen allemaal in dezelfde richting: anonimisering wordt steeds moeilijker, juridische categorisering zal steeds moeilijker worden en de status van gegevens zal steeds meer een effect zijn van de inspanningen van de verwerkingsverantwoordelijke. Er zijn wel meerdere scenario's gevonden voor hoe het wettelijk regime zou kunnen reageren op de toegenomen beschikbaarheid van open data en de algemene toegankelijkheid van dataverwerkingstechnologieën. Uit de suggesties kunnen vijf strategieën worden afgeleid, die als volgt kunnen worden samengevat:

1. **Het gegevensbeschermingskader laten zoals het is:** Het gegevensbeschermingskader wordt beschouwd als een perfect evenwicht tussen de beschermende grondgedachte en de bevorderende grondgedachte, tussen de keuze voor een categorale en een contextuele reguleringsbenadering en tussen het overlaten van het reguleringsprerogatief aan de wetgever en tegelijkertijd de rechterlijke en uitvoerende autoriteiten in staat te stellen concepten en regels in de praktijk te verfijnen, met het oog op specifieke contexten en situaties. Hoewel kan worden gezegd dat de technologische praktijk afwijkt van het reguleringsregime en dat deze afwijking in de loop der jaren heel goed steeds groter zou kunnen worden, betekent dit niet dat de regels moeten veranderen. Er moet veeleer meer worden geïnvesteerd om ervoor te zorgen dat de praktijk in overeenstemming blijft met de regels. Voor zover de verwerking van niet-persoonsgegevens een belangrijke impact heeft, wordt deze al gedekt door de AVG wanneer besluiten worden genomen waarin een individu wordt geïdentificeerd of op een persoonlijke manier wordt getroffen, of door artikel 8 EVRM, wanneer beleid van invloed is op het zeer brede begrip van privéleven. Het EHRM is bereid geweest om een regime te ontwikkelen voor het verzamelen van metadata, om een reguleringslacune te voorkomen, en heeft claims geaccepteerd waarin de eiser geen persoonlijk nadeel heeft geleden, maar heeft zich gericht op de maatschappelijke effecten van grootschalige gegevensverwerking. De wetgeving inzake gegevensbescherming hoeft bovendien niet alle problemen van de datagedreven omgeving op

te lossen.

2. **Handhaving van het gegevensbeschermingskader en investeren in nauwkeurigere definities:** De hoofdlijnen en contouren van het huidige reguleringsstelsel worden geschikt geacht voor de 21e eeuw, terwijl de belangrijkste uitdaging op het gebied van regelgeving de behoefte is aan meer duidelijkheid over de definities van de verschillende gegevenscategorieën, de grenzen tussen verschillende categorieën en de regulering van dat soort gegevens. In dit scenario zijn verschillende reguleringsalternatieven mogelijk, zoals het uitvaardigen van meer richtlijnen en het invoeren van een bewijslast voor de verwerkingsverantwoordelijke om aan te tonen dat gegevens anoniem en/of versleuteld zijn. Om meer duidelijkheid te scheppen over het onderscheid tussen niet-persoonsgegevens en persoonsgegevens, zouden de contextuele elementen in de definitie van persoonsgegevens en in de beschrijving van anonimisering kunnen worden verwijderd. Dit zou de vraag of persoonsgegevens worden verwerkt en of het kader voor gegevensbescherming van toepassing is, de-contextualiseren. Ook kan de categorie pseudonieme gegevens worden weggelaten. Deze categorie wordt zowel bekritiseerd vanwege zijn vaagheid als omdat het één privacybeschermende techniek voorrang geeft boven anderen, waarvoor geen duidelijke verklaring bestaat. Ten slotte kan worden overwogen om de lijst met gevoelige persoonsgegevens uit te breiden. Mogelijke aanvullende categorieën die in dit onderzoek zijn geïdentificeerd, zijn onder meer financiële en sociaaleconomische gegevens, gegevens over kinderen, locatiegegevens en metagegevens.
3. **Het gegevensbeschermingskader behouden en investeren in meer contextualiteit:** Als de belangrijkste uitdaging op het gebied van regelgeving wordt beschouwd het gebrek aan contextualiteit en aanpasbaarheid van het huidige reguleringsstelsel. Tijdens dit onderzoek zijn er verschillende reguleringsalternatieven naar voren gekomen, zoals, maar niet beperkt tot, de toevoeging van het contextualiteitsbeginsel aan de lijst van artikel 5 AVG, waarbij de verwerkingsverantwoordelijke wordt verplicht elk principe, elke verplichting en elk vereiste onder het gegevensbeschermingsregime in overweging te nemen in het kader van de context waarin de verwerking plaatsvindt. Als alternatief kan worden overwogen om de lijst van gevoelige gegevens te herformuleren zoals deze oorspronkelijk was, namelijk als voorbeelden in plaats van een uitputtende lijst, of om een restcategorie op te nemen, vergelijkbaar met artikel 14 EVRM. Pseudonieme gegevens zouden een meer prominente plaats kunnen krijgen als tussencategorie tussen niet-persoonsgegevens en persoonsgegevens.
4. **Herziening van het gegevensbeschermingskader met gebruikmaking van duidelijk gedefinieerde gegevenscategorieën:** Strategie 4 is vergelijkbaar met strategie 2, maar in dit scenario is een fundamentele herziening van het huidige reguleringskader noodzakelijk. In dit scenario wordt aangenomen dat het nog steeds mogelijk is om met gegevenscategorieën te werken, zelfs de huidige categorieën, maar in het licht van de technologische ontwikkelingen moet het reguleringsregime dat erop wordt toegepast worden heroverwogen. Een aantal reguleringsalternatieven zou kunnen worden overwogen, zoals het aannemen van een AVG-light regime voor niet-persoonsgegevens; dit zou bijvoorbeeld kunnen betekenen dat alle gegevensverwerkingen in overeenstemming moeten zijn met de beginselen van artikel 5 AVG. Ook zou, in het licht van een beschermend regime voor niet-persoonsgegevens, kunnen worden overwogen om het gegevensverwerkingsregime te structureren rond stadia van gegevensverwerking: het verzamelen en opslaan van gegevens, het analyseren van gegevens en het gebruiken van gegevens of de uitkomsten van gegevensanalyse. Het huidige reguleringsregime richt zich vrijwel uitsluitend op het moment dat gegevens worden verzameld en opgeslagen. Er zijn vrijwel geen regels voor de analyse van gegevens en voor het gebruik van gegevens, misschien met uitzondering van één bepaling over het verbod op geautomatiseerde besluitvorming. Dit wordt als problematisch ervaren omdat de kern van de meeste hedendaagse verwerkingen uit het analyseren van gegevens bestaat. Voor de inkadering van de analyse van gegevens zou inspiratie kunnen worden gezocht in de regels die gelden voor statistische bureaus.
5. **Herziening van het gegevensbeschermingskader, waarbij duidelijk gedefinieerde**

gegevenscategorieën worden verwijderd: Strategie 5 is vergelijkbaar met strategie 3, maar in dit scenario is een fundamentele herziening van het huidige reguleringskader noodzakelijk. In dit scenario is het eenvoudigweg onmogelijk om met verschillende gegevensdefinities te werken en aan elk van deze verschillende niveaus van wettelijke bescherming te koppelen. In plaats daarvan moet een volledig contextuele benadering worden gevolgd, die volledig afhankelijk is van een analyse van geval tot geval van de potentiële schade die het gevolg is van een bepaalde verwerking. Dergelijke schade kan verband houden met individuele belangen en/of maatschappelijke belangen. De meeste van de huidige verplichtingen en vereisten zouden intact kunnen blijven, maar ze zouden afhankelijk worden gemaakt van het risiconiveau. De AVG zou in wezen kunnen worden teruggebracht tot een eenvoudige set regels, namelijk een lijst van principes en verplichtingen voor verwerkingsverantwoordelijken die nu in de verordening staan en daarbij specificeren dat deze op hen van toepassing zijn, rekening houdend met de stand van de techniek, de kosten van uitvoering en de aard, omvang, context en doeleinden van de verwerking, de aard van de gegevens alsmede het risico en de ernst voor de individuele en/of maatschappelijke belangen.

Figuur: Schaal van een volledig categorale benadering (optie 4) naar een volledig contextuele benadering (optie 5)

10. Antwoorden op de onderzoeksvragen

1. Welke middelen zijn er om (anonieme) data terug te koppelen naar individuen en in hoeverre speelt de beschikbaarheid van andere (bijvoorbeeld open source) data een rol?

Er zijn veel middelen beschikbaar om gegevens terug te koppelen naar individuen. Dit onderzoek is niet tot een volledige en uitputtende lijst van mogelijkheden gekomen, maar heeft een aantal gangbare middelen besproken om dit te doen. Voorbeelden zijn databasereconstructieaanvallen (waardoor een geaggregeerde database opnieuw wordt geïdentificeerd), samenstelling (waardoor twee of meer geanonimiseerde datasets samengevoegd kunnen worden tot (gevoelige) persoonsgegevens) en verschillende de-anonimiseringstechnologieën. Uit geanonimiseerde datasets kan informatie worden afgeleid over personen die in eerste instantie niet in de dataset zaten en geaggregeerde data kunnen in het bijzonder worden gebruikt voor besluitvormingsprocessen die een significant effect kunnen hebben op burgers in het algemeen en specifieke groepen in bijzonder. Als dat laatste het geval is, kunnen die

gegevens kwalificeren als persoonsgegevens.

Open data speelt hierbij een belangrijke rol, zozeer zelfs dat veel experts erop wijzen dat het weliswaar mogelijk is om een geïsoleerde dataset te de-individualiseren, maar omdat het mogelijk is om deze te combineren met andere online vrij beschikbare data het nooit kan worden uitgesloten en het integendeel steeds waarschijnlijker zal worden, dat een geanonimiseerde dataset op termijn door een of andere partij wordt ge-deanonimiseerd. Geaggregeerde gegevens kunnen, wanneer ze beschikbaar worden gesteld, worden gebruikt voor besluitvorming die gevolgen heeft voor specifieke geïdentificeerde of niet-geïdentificeerde burgers. Hoe gegevens zullen worden gebruikt, kan vooraf niet met zekerheid worden gecontroleerd of ingeschat. Echter, de kans dat wanneer data online beschikbaar wordt gesteld, deze door een partij worden gebruikt op manieren die effect hebben op concrete individuen, groepen of de samenleving als geheel, wordt steeds groter.

2. Welke (technische) ontwikkelingen worden de komende jaren verwacht met betrekking tot de middelen om gegevens (al dan niet opzettelijk) terug te koppelen aan personen?

Het zal steeds moeilijker worden om de (juridische) anonimiteit van datasets te waarborgen. Experts die voor dit onderzoek zijn geïnterviewd, betwijfelen nu al of het mogelijk is om aan de wettelijke criteria voor anonimiteit te voldoen. Terwijl het wettelijke regime anonimiteit als een binair vraagstuk beschouwt, zien de meeste technische experts het als een schaal. De meeste technologieën en tegen-technologieën zijn verwikkeld in een kat-en-muisspel. Dit wordt ook verondersteld het geval te zijn voor de toekomst van onder meer anonimiserings- en de-anonimiseringstechnieken, aggregatie- en inferentietechnieken en voor encryptie en decryptie. De meest fundamentele verschuiving is de algemene toegankelijkheid van dergelijke technologieën. Dit betekent dat, vooral wanneer gegevens online beschikbaar worden gesteld, het steeds waarschijnlijker wordt dat er wereldwijd enkele partijen zullen zijn die geavanceerde technologieën zullen gebruiken om gegevens te ontsleutelen, opnieuw te identificeren of te de-anonimiseren en de nodige tijd, energie en moeite zullen investeren om dit te doen. Een belangrijke ontwikkeling op het gebied van encryptie is quantum computing.

Quantum computing heeft bepaalde kenmerken die zijn afgeleid van de kwantummechanica en die het mogelijk maken om complexe factorisatieproblemen op te lossen waarmee traditionele computers worstelen. In plaats van met bits te werken, werken kwantumcomputers met kwantumbits of qubits. Qubits kunnen tegelijkertijd een waarde van 0 of 1 aannemen, in tegenstelling tot traditionele bits, die enkel een toestand van ofwel 0 ofwel 1 hebben. Hierdoor kunnen kwantumcomputers meerdere parallelle berekeningen uitvoeren waarvoor conventionele computers niet geschikt zijn. Als gevolg hiervan kan quantum computing mogelijke alle huidige vormen van cryptografie kraken, net zoals de huidige technieken de met Data Encryption Standard (DES) versleutelde berichten van 40 jaar geleden kunnen ontsleutelen.

3. Welke actuele en voorzienbare technische ontwikkelingen kunnen worden gebruikt voor het anonimiseren of pseudonimiseren van persoonsgegevens en welke factoren zijn daarbij bepalend?

Er bestaan verschillende technieken voor zowel anonimisering als pseudonimisering. Voorbeelden van anonimiseringstechnieken omvatten, maar zijn niet beperkt tot: het maskeren en gebruiken van synthetische gegevens. Er zijn verschillende factoren die bepalend zijn, maar veel hangt af van de vraag of er een technische of een juridische benadering wordt gekozen. Ook zijn in de technische literatuur verschillende soorten anonimiteit naar voren gebracht, elk met hun eigen nadruk op verschillende factoren, met als belangrijksten: k-anonimiteit, l-diversiteit, t-nabijheid en ε-differentiële privacy.

Voor aggregatie kan onderscheid worden gemaakt tussen onder meer aggregatie op basis van derden, aggregatie op basis van dataverstoring en aggregatie op basis van cryptografie. Elk daarvan onderstreept verschillende factoren die bepalend worden geacht. Misschien wel de belangrijkste techniek voor het

aggregeren van gegevens, vooral in het licht van het vrijgeven van gegevens, is SDC. Er is geen vaste standaard voor SDC; elke organisatie kan zijn eigen factoren, normen en drempels hanteren, rekening houdend met de dataset, de waarde ervan en mogelijke privacy risico's.

Er bestaan verschillende pseudonimiseringstechnieken, de belangrijksten voor de doeleinden van dit onderzoek: hashing, key hashing, salt hashing en pepper hashing. Encryptie wordt juridisch gezien beschouwd als een deelverzameling van pseudonimisering. Er bestaan verschillende encryptietechnieken, de belangrijksten: symmetrische encryptie, asymmetrische encryptie, homomorfe encryptie en multiparty-computation (wat meer is dan enkel een encryptietechniek). De laatste is een techniek die zich bezighoudt met protocollen waarmee een reeks partijen gezamenlijk een functie van hun invoer of identificatiegegevens kan berekenen, terwijl wordt vermeden dat iets anders wordt onthuld dan de uitvoer van die functie.

4. Welke technische ontwikkelingen op het gebied van anonimisering en pseudonimisering van persoonsgegevens zijn de komende jaren te verwachten?

De meeste geïnterviewde experts en de voor dit onderzoek geëvalueerde literatuur verwachten geen technologische revolutie op het gebied van anonimisering en pseudonimisering, maar verwachten dat het kat-en-muisspel de komende jaren zal doorgaan. Door de steeds grotere beschikbaarheid van gegevens en de algemene toegankelijkheid van technologieën kan het echter nog moeilijker worden om tot anonieme of pseudonieme gegevens te komen. Quantum computing kan, zoals gezegd, een belangrijke impact hebben op encryptie. Daarnaast zal ‘deep learning’ naar verwachting de komende jaren nog meer bekendheid krijgen. Beide technologieën kunnen een nadelig effect hebben op privacy, maar ze kunnen ook in het voordeel van privacy worden ingezet. Post-kwantumversleuteling wordt als veel veiliger beschouwd dan de huidige vormen van versleuteling, en momenteel worden reeds deep privacy-tools (privacy-tools op basis van deep learning-modellen) ontwikkeld.

5. Wat kan er vanuit het juridisch en technisch perspectief gezegd worden over de interpretatie van het begrip ‘alle middelen waarvan redelijkerwijs valt te verwachten dat zij worden gebruikt’? Welke middelen zijn redelijkerwijs in te zetten en welke factoren spelen daarbij een rol?

Vanuit juridisch oogpunt hebben zowel het HvJ EU als de Artikel 29 Werkgroep keer op keer benadrukt dat de beoordeling van welke middelen redelijkerwijs kunnen worden geacht te worden gebruikt per geval moet worden gemaakt, rekening houdend met alle relevante omstandigheden van het geval en met oog op verschillende relevante, maar niet op zich bepalende factoren, zoals de kosten en de tijd die nodig zijn voor identificatie, de beschikbare technologie op het moment van de verwerking en technologische ontwikkelingen. Hoewel dit op zichzelf objectieve criteria zijn, hangt de interpretatie ervan af van de context. Dus hoewel het onderscheid tussen niet-persoonsgegevens en persoonsgegevens juridisch binair en absoluut is, zijn de criteria om te bepalen of gegevens anoniem zijn zeer contextueel.

Vanuit technisch perspectief is de contextuele benadering het meest voor de hand liggend. De meeste technische experts geloven niet in absolute of volledige anonimiteit, maar wijzen eerder op een schaal van hoe moeilijk het is om een database te de-anonimiseren of opnieuw te identificeren. Omdat de technologische mogelijkheden voor de-anonimisering evolueren, moet een beoordeling van de technische normen om gegevens te anonimiseren mogelijk continue of periodiek gebeuren. Een zwart-wit onderscheid tussen anonieme en niet-anonieme gegevens ligt in dit verband niet voor de hand; vanuit technisch oogpunt zou het eerder passender kunnen zijn om te werken met een schaal waarbij hoe anoniemer gegevens zijn, hoe minder (strenge) gegevensbeschermingsnormen van toepassing zijn. Er is geen uitputtende lijst van factoren vanuit een technologisch perspectief waarmee rekening moet worden gehouden om de redelijk waarschijnlijke middelen te bepalen (een juridisch begrip dat in de meeste technologische discussies niet voorkomt).

6. Hoe verhoudt het antwoord op vraag 5 zich tot ontwikkelingen in huidige en toekomstige anonimiserings- en pseudonimiseringstechnieken?

De algemene beschikbaarheid van open data en de algemene toegankelijkheid van datatechnologieën zullen een drievoudig effect hebben op de mogelijkheden om anonimisering en pseudonimisering te realiseren.

Ten eerste is de aard van de data in Big Data-processen niet stabiel, maar volatiel. Een dataset met gewone persoonsgegevens kan worden gekoppeld aan en verrijkt met een andere dataset om gevoelige gegevens af te leiden; de gegevens kunnen vervolgens worden samengevoegd of ontdaan van identificatiegegevens en niet-persoonlijk worden, zoals geaggregeerde of anonieme gegevens; vervolgens kunnen de gegevens worden gedeanonimiseerd of geïntegreerd in een andere dataset om opnieuw persoonsgegevens te creëren. Dit alles kan in een fractie van een seconde gebeuren. De vraag is daarom of het zin heeft om met afgebakende categorieën te werken als dezelfde 'datum' of dataset letterlijk van de ene seconde op de andere in een andere categorie kan vallen en de volgende seconde in weer een andere.

Ten tweede wordt het als gevolg van het voorgaande steeds moeilijker om de status van gegevens precies te bepalen. Om de huidige status van een datum of dataset te bepalen, moet rekening worden gehouden met de verwachte toekomstige status van de gegevens. Gezien de algemene toegankelijkheid van technologieën en de minimale investering die nodig is, wordt het steeds waarschijnlijker dat wanneer een database wordt gedeeld of anderszins beschikbaar wordt gesteld, er een partij is die deze combineert met andere data, deze verrijkt met data van internet geschraapt of samenvoegt in een bestaande dataset, maar ook dat er andere partijen zijn die dat niet willen. De juridische categorie waartoe de gegevens behoren, is dus niet langer een kwaliteit van de gegevens zelf, maar een product van de inspanningen en investeringen van een verwerkingsverantwoordelijke. Bijgevolg is het de vraag of anonimisering of pseudonimisering kan worden bereikt in een context waarin het bepalen van de status van gegevens nauwelijks haalbaar is.

Ten derde zijn moderne gegevensverwerkingen in toenemende mate gebaseerd op geaggregeerde gegevens, die ook zeer grote individuele en sociale gevolgen kunnen hebben. Het profileren van doelgroepen in plaats van individuen wordt een gangbare verwerkingshandeling in de informatiemaatschappij. De gevolgen van deze activiteiten kunnen negatief zijn voor de groep, zonder dat de schade direct te relateren is aan individuen. Het idee dat hoe gevoeliger de gegevens zijn en hoe directer ze aan een persoon kunnen worden gekoppeld, des te strikter de verwerking ervan moet worden gereguleerd, kan daarom in twijfel worden getrokken. Daarnaast is het de vraag of de focus op de identificeerbaarheid van een individu (natuurlijke persoon) en vervolgens de noties van anonimisering en pseudonimisering die daarop zijn gebaseerd, te handhaven zijn in de 21e eeuw.

7. Wanneer is het redelijk om te zeggen dat gegevens niet meer terug kunnen worden gekoppeld aan een persoon en dat de dataset waarvan ze deel uitmaken als anoniem kan worden beschouwd?

Hoewel er vanuit juridisch oogpunt een verschil is tussen niet-persoonsgegevens en persoonsgegevens, valt dit onderscheid vanuit technisch oogpunt uiteen in ten minste drie relevante subcategorieën:

1. de situatie waarin gegevens nooit persoonlijk waren, maar wel zouden kunnen zijn, zoals wanneer klimaatdata worden gebruikt om beslissingen te nemen over de verzekering van individuele boeren;
2. de situatie waarin gegevens persoonlijk waren, maar de identifiers zijn gestript of gegevens zijn geanonimiseerd op een zodanige manier dat de betrokkene niet kan worden geïdentificeerd of identificeerbaar is. Hierbij bestaat het gevaar dat gegevens opnieuw worden gereïdentificeerd of gedeanonimiseerd;

3. de situatie waarin gegevens worden geaggregeerd. Hierbij bestaat zowel het gevaar dat gegevens kunnen worden gedeaggregeerd, dat de combinatie van twee geaggregeerde datasets persoonsgegevens kunnen opleveren, en dat geaggregeerde gegevens kunnen worden gebruikt om beslissingen te nemen die van invloed zijn op individuele betrokkenen of om hen eruit te pikken, zonder hun identiteit te kennen.

Voor elk van die scenario's zijn er verschillende bedreigingen. Vanuit het technologische domein is duidelijk dat het bijna nooit aannemelijk is dat data niet meer terug te koppelen zijn aan een individu. Er zijn altijd risico's voor de-anonimisering, er zijn altijd mogelijkheden tot gegevenssamenstelling en het kan nooit worden uitgesloten dat gegevens worden gebruikt om niet-geïdentificeerde individuen te onderscheiden of om beslisbomen te ontwikkelen die een impact hebben op groepen en/of individuen. Daardoor is het steeds moeilijker te bevestigen dat data niet meer terug te koppelen zijn aan een individu en dat de dataset waarvan ze deel uitmaken als anoniem kan worden beschouwd.

8. In hoeverre is de toets op indirecte identificeerbaarheid objectiveerbaar?

Er zijn maar weinig aanwijzingen gevonden om de toets op indirecte identificeerbaarheid meer objectiveerbaar te maken. Het is belangrijk om te onderstrepen dat het objectiveerbaar maken van de test niet het doel was van de EU-regelgever. Integendeel, de huidige open, contextuele en fluïde reeks criteria kreeg de voorkeur boven de meer beperkende criteria die werden overwogen en verworpen. Zo bevatte het oorspronkelijke voorstel voor de richtlijn gegevensbescherming niet het begrip anonimiteit, maar veeleer dat van ‘depersonalisatie’, dat werd opgevat als het zodanig wijzigen van informatie dat het niet langer aan een specifiek individu kon worden gekoppeld. In de memorie van toelichting werd bepaald dat ‘een gegeven kan worden beschouwd als gedepersonaliseerd, zelfs als het theoretisch zou kunnen worden gerepersonaliseerd met behulp van onevenredige technische en financiële middelen’. Tegelijkertijd definieerde de toelichting depersonalisatie als "het zodanig wijzigen van persoonsgegevens dat de informatie die ze bevatten niet langer in verband kan worden gebracht met een specifieke persoon of een persoon die kan worden bepaald, behalve tegen de prijs van een buitensporige inspanning.” Overmatige inspanning is nog steeds contextueel, maar minder dan “alle middelen waarvan redelijkerwijs valt te verwachten dat zij worden gebruik”; ook is de drempel duidelijk anders.

Er zijn in dit onderzoek weinig aanwijzingen gevonden om de toetsing van indirecte identificeerbaarheid objectiever te maken dan het schrappen van het begrip 'identificeerbaarheid', dat oorspronkelijk geen deel uitmaakte van de definitie van persoonsgegevens onder de gegevensbeschermingsregimes van vóór 1995, of het beperken van de lijst van factoren die moeten worden opgenomen om te bepalen welke middelen redelijkerwijs moeten worden gebruikt. Misschien is de enige concrete suggestie die werd geïdentificeerd, het stellen van een tijdsbeperking of een horizon voor de evaluatie van de middelen die redelijkerwijs kunnen worden gebruikt. Het is bijna altijd zeer waarschijnlijk dat over 20 jaar gegevens die nu anoniem zijn, kunnen worden gedeanonimiseerd. Onder het huidige wettelijke regime moet, wanneer gegevens zo lang worden bewaard of als ze openbaar worden gemaakt, rekening worden gehouden met dergelijke middelen die redelijkerwijs kunnen worden gebruikt bij het bepalen of het gegevensbeschermingsregime van toepassing is, terwijl het vrijwel onmogelijk is te voorzien hoe het technologische landschap en de beschikbaarheid van data zich de komende 20 jaar zal ontwikkelen.

9. In hoeverre en in welke gevallen kan er sprake zijn van onderregulering wanneer gegevens niet meer door middel van anonimisering aan personen worden gekoppeld en dus niet onder de AVG vallen?
10. In welke mate en in welke gevallen kan er sprake zijn van overregulering wanneer steeds meer gegevens eenvoudig aan individuen kunnen worden gekoppeld door middel van nieuwe technieken (het ongedaan maken van maatregelen van anonimisering en pseudonimisering)?

Het beantwoorden van de vragen 9 en 10 hangt af van wat wordt beschouwd als de

reguleringsdoelstelling van het gegevensbeschermingsregime: moet het gegevensbeschermingskader worden beschouwd vanuit een beschermend perspectief of vanuit het perspectief van het faciliteren van gegevensverwerking binnen een vastgesteld kader, of als een combinatie tussen beide? Moet het beschermingsdoel worden opgevat als het voornamelijk bieden van bescherming aan individuele belangen of (ook) aan groeps- en maatschappelijke belangen? Moet het gegevensbeschermingsregime worden opgevat als het stellen van beperkingen voor gegevensverwerking of als het bieden van een kader voor het gebruik en het delen van gegevens? Is de beschermende grondgedachte het best gediend door beperkingen, of kan er soms meer gegevensverwerking nodig zijn om de belangen van individuen en/of de samenleving zo goed mogelijk te dienen? Is de grondgedachte van het faciliteren van gegevensgebruik het best gediend met een open en contextueel kader of met het stellen van strikte en duidelijke regels waarbinnen gegevensverwerking als legitiem wordt beschouwd? Dit onderzoek heeft niet ten doel om op deze vragen een definitief antwoord te geven; wel is duidelijk dat afhankelijk van de antwoorden op deze vragen verschillende lacunes in de regelgeving en gevaren voor over- en/of onderregulering zullen worden gevonden.

Of er bijvoorbeeld sprake is van onderregulering omdat ‘persoonsgegevens’ alleen gekoppeld zijn aan de identificeerbaarheid van natuurlijke personen en omdat het gegevensbeschermingskader primair verwijst naar de belangen van de betrokkene, hangt af van wat als het kerndoel van het gegevensbeschermingskader wordt gezien. Als wordt aangenomen dat het gegevensbeschermingskader bescherming biedt of zou moeten bieden aan meer algemene, groeps- of maatschappelijke belangen, dan kan er zeker sprake zijn van onderregulering omdat de verwerking van geaggregeerde en anonieme gegevens niet onder het huidige regime valt. Of de trend van rechters en adviesorganen om de reikwijdte van persoonsgegevens en de materiële reikwijdte van het gegevensbeschermingskader uit te breiden tot overregulering leidt, hangt af van de vraag of de nadruk wordt gelegd op de beschermende grondgedachte van het gegevensbeschermingskader, in welk geval er geen sprake zou zijn van overregulering, maar het juist toe valt te juichen dat de reikwijdte steeds wordt uitgebreid, of dat de nadruk wordt gelegd op het faciliterende doel, in welk geval een te grote reikwijdte van het gegevensbeschermingsregime sneller als overregulering kan worden beschouwd.

11. Hoe zullen de huidige en toekomstige technische ontwikkelingen de komende periode van invloed zijn op de AVG en de rechtsbescherming van gegevens in brede zin?

Het is duidelijk dat de technologische ontwikkelingen en de algemene beschikbaarheid van data nu en in de toekomst tot gevolg hebben dat anonimisering steeds moeilijker wordt. De status van data wordt steeds volatieler en wordt steeds minder een kenmerk van data en datasets zelf en meer en meer een effect van de inspanningen van de verwerkingsverantwoordelijke. De juridische categorieën zullen steeds meer fluïde en minder stabiel worden en één database kan juridisch verschillend worden beoordeeld ten aanzien van verschillende partijen die daar toegang toe hebben. Een database die op zichzelf alleen niet-persoonsgegevens bevat kan worden omgezet in persoonsgegevens door deze het volgende moment te combineren met een andere database, kan vervolgens worden gebruikt om gevoelige persoonsgegevens af te leiden, om het moment daarop weer te worden geaggregeerd en geanonimiseerd. Gezien deze trends en gezien de begrippen ‘identificeerbaarheid’ en ‘alle middelen waarvan redelijkerwijs valt te verwachten dat zij worden gebruik’, zullen steeds meer gegevens, zo niet alle, onder het gegevensbeschermingskader vallen.

In dit onderzoek zijn geen verschillende scenario's gevonden voor hoe de technologische wereld en de beschikbaarheid van open data zich in de loop van de tijd zullen ontwikkelen: literatuur, geïnterviewde experts en experts die zijn uitgenodigd voor de workshop die voor dit onderzoek is gehouden, wijzen allemaal in dezelfde richting. Er zijn echter meerdere scenario's gevonden voor hoe het wettelijk regime zou kunnen reageren op de toegenomen beschikbaarheid van open data en de algemene toegankelijkheid van technologie. Uit de suggesties zijn vijf globale strategieën afgeleid: het huidige gegevensbeschermingskader intact laten, focussen op duidelijkere gegevenscategorieën, meer nadruk

leggen op contextualiteit, gebruik maken van verschillende gegevenscategorieën en daaraan gekoppelde reguleringsregimes, of focussen op een volledig contextueel gegevensbeschermingskader.

Chapter 1: Study design

1.1 Introduction

This chapter will introduce the background of this study (section 1.2), the research questions that guide the study (sections 1.3) and the core methodologies used (section 1.4). Then, it will provide a reader's guide through the main themes and topics of the report (section 1.5) and finally, it presents an overview of the report and describes the content of the various chapters (section 1.6).

1.2 Problem statement

In Europe, there is an ambiguous approach to the right to data protection and the scope of 'personal data' in particular.

On the one hand, the European Union is set on maintaining a strict separation between personal and non-personal data, as well as other categories. While personal data are protected under arguably the world's strictest regime (the General Data Protection Regulation),[1] non-personal data are almost free from regulation, or, to put it more precise, the EU has adopted a Regulation on non-personal data in which it dissuades public and private sector organisations alike from adopting any restrictions on or barriers to the free flow of non-personal data.[2] This choice stands in a broader tradition within the EU for opting for separate, demarcated types of data that each have their own level of protection. In addition to the distinction between non-personal and personal, the GDPR differentiates between anonymous and identifying, directly identifying data and pseudonyms, and 'ordinary' personal data and 'sensitive' personal data. Numerous adjoining legal instruments have their own data concepts, each of which has been assigned its own scope and level of protection. Examples are the proposed e-Privacy Regulation,[3] which makes a distinction between, among others, 'electronic communications data', 'electronic communications content', 'electronic communications metadata' and 'location data', and the proposed AI Act, which differentiates between 'training data', 'testing data', 'input data' and 'biometric data'. The presumption that guides EU regulation is that data can be distinguished and demarcated reasonably well and that separate regimes of protection can be attached to them.

On the other hand, the concept of 'personal data' has been extended in the various data protection instruments adopted over the decades. In case law, courts have also given a broad interpretation to the definition. Thus, scientific opinions, open access data, dynamic IP addresses, minutes with draft decisions about persons, registration of working hours by employees, and metadata may all fall under its scope. The various advisory bodies, such as the Article 29 Working Party (WP29), have propagated a broad approach to the material scope of data protection regimes too. The reason is that over time, more and more data can be used to identify a person or make decisions that affect a person. It is also relatively easy to combine various non-sensitive data points and, through predictive analysis, infer sensitive personal data, for example, saying something about a person's prospective health. What counts as non-personal, personal, or sensitive personal data has become increasingly difficult to establish and more and more fluid over time and will continue to be. To provide for a high level of protection, the various concepts and scopes have been widened over the years.

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).
[2] Regulation (EU) 2018/1807 of the European parliament and of the council of 14 November 2018 on a framework for the free flow of non-personal data in the European union.
[3] Proposal for a Regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC.

The first approach, that of having several strictly separated types of data, each with their own scope of protection, is increasingly criticized. Broadly speaking, three arguments can be put forward.

First, it is argued that working with well-defined and delimited definitions of different types of data only works if the status of data is relatively stable, if a 'datum' falls into one category in a relatively stable way. This is exactly what is increasingly less so. The nature of the data in Big Data processes is not stable, but volatile. A dataset containing ordinary personal data can be linked to and enriched with another dataset so as to derive sensitive data; the data can then be aggregated or stripped of identifiers and become non-personal, aggregated, or anonymous data; subsequently, the data can be deanonymized or integrated into another dataset in order to create personal data. All this can happen in a split second. The question is, therefore, whether it makes sense to work with well-defined categories if the same 'datum' or dataset can literally fall into a different category from one second to the next and into still another the very next second.

Second, it is also increasingly difficult to determine the status of data precisely. As the Working Party 29 already stated: 'the assessment of whether the data allow identification of an individual, and whether the information can be considered as anonymous or not depends on the circumstances, and a case-by-case analysis should be carried out with particular reference to the extent that the means are likely reasonably to be used for identification'.[4] This refers to the phrase in the GDPR, holding that in order to determine whether a datum is to be considered 'personal', account should be had of the means that can reasonably be expected to be used for identification. Therefore, in order to determine the current status of a datum or dataset, the expected future status of the data must be taken into account. Given the general availability of technologies and the minimal investment required, it is increasingly likely that when a database is shared or otherwise made available, there will be a party who will combine it with other data, enrich it with data scraped from the internet, or merge it into an existing dataset. It is thus increasingly likely that if an anonymised dataset is made public, there will be a party that will deanonymize it or combine it with other data to create personal profiles; that if a set of personal data is shared, there will be a party that will use that data to create a dataset containing sensitive personal data; and so on. On the other hand, there will be other parties who have access to that data but will not engage in such activities; parties who will not use the data, use it as it is provided, or even de-identify a database containing personal data. Who will do what is not clear in advance. The legal category to which the data belongs is therefore no longer a quality of the data itself, but a product of a data controller's efforts and investments.

Third, the question is whether the distinction made between different categories of data is still relevant. The underlying rationale is that the processing of personal data has an effect on natural persons, while the processing of non-personal data does not and that the processing of sensitive personal data may have very significant consequences (greater than the processing of 'ordinary' personal data normally has), so that the latter are subject to the most stringent regime, personal data fall under the 'normal' protection regime, and the processing of non-personal data is not subject to any restrictions. Pseudonymisation does not ensure the full protection of individuals, but it does greatly reduce the number of people and organisations that can link data to specific individuals, which is why pseudonymous data are put in an intermediate category of protection. The question is to what extent this rationale is still tenable in the 21st century. Not only can information about the content of communication be distilled from metadata, can identifying data be inferred by combining two datasets holding no personal data, etc., modern data processing on the basis of aggregated data, for example, can also have very large individual and social consequences. Profiling, by definition, targets groups rather than individuals. The consequences of profiling can be negative for groups, without the damage being directly relatable to individuals, such as when the police, using predictive policing, decides to patrol certain neighbourhoods more often than others. The possible arrests made in these neighbourhoods may all be justified in and by themselves,

[4] Article 29 Working Party, Opinion 4/2007 on the concept of personal data, 01248/07/EN WP 136, 20 June 2017, p. 21.

while the general problem of stigmatisation of deprived neighbourhoods and blind spots on the part of the police with regard to 'better' neighbourhoods may be significant. The same applies to profiles used in smart cities. The idea that the more sensitive the data are and the more directly they can be linked to a person, the more strictly its processing should be regulated, can therefore be questioned.

On the other hand, the second regulatory approach, that is to continue stretching the notion of personal data and of sensitive personal data, so that more and more data fall under those categories, is also criticised, as it would effectively make data protection law applicable to virtually all processes in an increasingly data-driven society. In addition, by accepting that more and more personal data may indirectly disclose sensitive personal data (e.g. the fact that two men live together and listen to certain types of music via Spotify combined may, under circumstances, be enough to derive predictive information about their sexual preferences), more and more data processing initiatives will be put under the strictest regulatory regime. This approach may stifle innovation, reduce economic growth, and block data processing initiatives that serve personal and societal interests.

1.3 Research questions

Given what has been described in the previous section, the research question for study is:

What effect do current and future technical developments with respect to the anonymisation, pseudonymisation, aggregation and identification of data have on the data protection framework and the protection afforded to the different types of data?

The sub-questions that help answer this research question are:

Identifiability of data

1. What means are available to link (anonymous) data back to individuals, and to what extent does the availability of other (e.g. open source) data play a role?
2. Which (technical) developments are expected in the coming years with regard to the means to (intentionally or unintentionally) link data back to persons?

Anonymisation and pseudonymisation of data

3. What current and foreseeable technical developments can be used for the anonymisation or pseudonymisation of personal data, and what factors are decisive in this respect?
4. What technical developments in the area of anonymisation and pseudonymisation of personal data are to be expected in the coming years?

Identifiability in relation to anonymisation and pseudonymisation and vice versa

5. What can be said, from a legal and technical perspective, about the interpretation of the concept of 'means reasonably likely to be used': what means can be considered reasonably likely to be used, and what factors play a role in this?
6. How does the answer to question 5 relate to developments in current and expected techniques to achieve anonymisation and pseudonymisation?
7. When is it reasonable to say that data can no longer be linked back to an individual and that the dataset of which they are part can be considered anonymous?
8. To what extent is the test for indirect identifiability objectifiable?

Consequences of identifiability and anonymisation and pseudonymisation

9. To what extent and in which cases can there be underregulation when data are no longer linked to individuals through anonymisation and therefore do not fall within the scope of the GDPR?
10. To what extent and in which cases can there be overregulation when more and more data can be easily linked to individuals through new techniques (undoing measures of anonymisation and pseudonymisation)?

Overarching analysis

11. How will the current and future technical developments affect the GDPR and legal protection in a broad sense in the coming period?

1.4 Methodology

The identifiability of data, which determines the line between personal and non-personal data, anonymous and non-anonymous data, and pseudonymised or not fully pseudonymised data, is dependent on the state of the art of technologies. The same argument applies to information inference which can be used to infer sensitive data from non-sensitive data. The literature on the technical aspects of identifying techniques and protective techniques has shown over the years that certain assumptions about anonymization do not hold, as individuals in datasets presumed to be anonymous could relatively easily be identified.[5]

Famous is the research 'Unique in the crowd', in which only four points in time and place were enough to trace 95 percent of the sample of 'anonymous data' of mobile phone use to unique persons.[6] Similarly, it is increasingly easy to identify individuals in aggregated data such as statistical data.[7] The availability and linkability of the large volumes and varieties of data available in the world only sketch part of the story, an analysis of identifying technologies on the one hand and privacy or data protection preserving technologies on the other hand is required to assess how to continue to regulate the different categories of data to offer meaningful protection.

The legal and technical dimensions thus need to be considered simultaneously. In addition to combining a technical and legal perspective, it is crucial to include empirical research which can show more clearly where in the practise of data processing the challenges lie vis a vis using different categories of data. In addition, the practice of data sharing and the general availability of data, for example, online in open access databases, is relevant for the possibility of de- and re-identification. That is why this report will look into three domains. The legal regulations will be studied, the technical developments and applications will be outlined, and the practical availability of data and use cases will be mapped.

Within the legal domain, a diversified approach will be taken. The challenges of maintaining data protection are an international problem not limited to a specific jurisdiction. It is, therefore, paramount to examine this problem from a broader perspective than just the EU perspective. In the EU, the GDPR applies. Over the past years, in other jurisdictions, there have been major developments in data protection legislation as well, such as in India[8], but also on the state level in the USA. The concepts of personal data, anonymous data, pseudonymous data, and sensitive data are concepts of the GDPR; in other jurisdictions outside of the EU, different approaches to types of data are taken. Even within the EU, there is also some margin for member states to further detail aspects of the GDPR, for example,

[5] See the AOL case for example: <https://www.youtube.com/watch?v=c-SOCGdPyNU>.
[6] De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. Scientific reports, 3(1), 1-5.
[7] Fluitt, A., Cohen, A., Altman, M., Nissim, K., Viljoen, S., & Wood, A. (2019). Data Protection's Composition Problem. Eur. Data Prot. L. Rev., 5, 285.
[8] Bailey, R., Parsheera, S., & Sane, R. (2020). Comments on the Report by the Committee of Experts on Non-Personal Data Governance Framework'. Available at SSRN 3724184.

when it comes to the regulation of safeguards for statistical and research data.[9] In addition, the European Court of Human Rights and its jurisprudence on the right to privacy contained in the European Convention on Human Rights plays an important role.

For this study, three methodological approaches are deployed.

(1) Doctrinal and legal analysis

In this report, four types of data distinctions are discussed in four chapters: the distinction between anonymous and personal data (chapter 2), the distinction between aggregated or statistical data and personal data (chapter 3), the distinction between pseudonymous and non-pseudonymous data (chapter 4) and the distinction between sensitive and non-sensitive personal data (chapter 5). For each chapter addressing a different category of data, the provisions that regulate that type of data on the EU and CoE level are studied, as well as their legislative history, policy documents or guidelines, such as Article 29 Working Party Opinions, and case law by the European Court of Human Rights (ECtHR) and the Court of Justice of the EU (CJEU).

(2) Literature review

In addition to the doctrinal research, two main literature studies were conducted.

2a. Descriptive literature

The first literature study concerned the body of technical literature on identifying techniques and privacy/data protection protective techniques. The purpose of that literature study was twofold: to identify the most important techniques on both sides of the coin, and to address the strengths and weaknesses of said techniques to allow for a critical reflection. The selected literature started from a quick international scan of the field of identification and anonymization techniques. Subsequently, prominent authors or developers were selected as well as techniques that showed either promising developments according to the technical field or are most known or widely used. These were selected based on, inter alia, citation and reference number, as well as those pointed out by interviewees. Then, for the critical discussion, not only the literature of the proponents or developers of said techniques was used but also literature that criticizes the use of such techniques and highlights challenges of weaknesses. Given the scope of the research, which is to ultimately link the technical state of the art back to the regulatory framework, it is not possible to consider every possible technique. Rather, the literature study focuses on prominent developments and applications to showcase what is possible in terms of identifying individuals. These were selected on the basis of the workshop held for this study.

2b. Normative literature

The second main literature review looked at a combination of legal literature that describes the challenges of each category of data, and that proposes new definitions, perspectives, or approaches to anonymous, pseudonymous personal data and other types of personal data while taking into account or referring to various identification techniques. This literature review was intended to bridge the gap between doctrinal research and the technical literature review, and to analyse the problems for each type of data that is addressed in this study. Included in this review are the prominent criticisms of data protection scholars on the relevant concepts and definitions of the GDPR, as well as reflections on identification and identifiability from leading scholars from other jurisdictions, such as the USA. This literature review supported the analysis provided at the end of each chapter and resulted in the overview of regulatory alternatives presented in section 6.6.

[9] GDPR, Articles 89(2) and (3).

(3) Qualitative research methods

The third research method deployed for this report is a qualitative research method. Two types of qualitative research methods are used for this study: interviews and a workshop. The interviews are intended to gather knowledge on specific domains. Through the interviews, experts are questioned on specific themes, developments, and niche issues. The workshop was used to facilitate discussions between experts with technical, legal, and policy background. Many of the questions that are central to this study arise from a mismatch between technical developments, their understanding by lawyers, their eventual reflection in legal frameworks, and the applications as used in practice.

3a. Interviews

The interviews were conducted with experts with different backgrounds and areas of expertise: experts on mainly one specific technique, experts with an overview of anonymization/pseudonymization techniques, experts from organisations that work with a lot of data in practise to discuss practical challenges. The interviews were used to get knowledge on the technological state of the art when it comes to identifying individuals or protecting the identity of individuals, as well as on practical challenges of applying the GDPR concepts to data. Therefore, ten people were interviewed for this study. For each interview, in preparation, the interviewers studied publications or policy documents written by the interviewee in question or by the organizational unit the interviewee works for. The expert interviews were semi-structured, thus, for each interview, a pool of questions was prepared by the interviewers inspired by the aforementioned documents, but follow-up questions were posed as well. The interviewees and topics of the interviews were selected both as a result of the first exploratory workshop as well as based on the literature study. The full interviews reports are presented in the annex to this report. The main findings, as per the distinction between the various categories of data, are provided in chapters 2-5.

3b. Workshop

In addition, a workshop was held at the beginning of the study. The workshop was intended to identify problems and mismatches between the legal and policy domain on the one hand and the technical and practical reality on the other. The full workshop report is presented in the annex to this report. The main findings, as per the distinction between the various categories of data, are provided in chapters 2-5.

1.5 Themes and topics discussed in this report

To answer sub-questions 1-8, several factors will be discussed in this report:

- The various legal concepts and the criteria that define and demarcate them;
- The availability of (open access) data and of data processing technologies; in this respect, the European Union's (EU) push for open data and re-use of data is relevant;
- The current and future technological means for anonymising and deanonymizing, aggregating and de-aggregating, pseudonymising and de-pseudonymising data, etc.;
- The impact of the evolving technological capabilities and expanding data landscape on the viability of current legal concepts and demarcations.

To answer sub-questions 9-11, several aspects are relevant:

- The regulatory objective of the data protection framework and the light in which the danger of both under- and overregulation should consequently be assessed
- The regulatory gaps that emerge from the disconnect between the legal and the technological realm

- The alternatives to the current legal framework can be gained from previous European legislation and legislative proposals, from literature, and from interviews

In answering questions 9-11 and in order to determine whether there is under- and/or over-regulation, it must be determined what the regulatory objective of the GDPR is and should be. Two matters need to be examined in this regard. On the one hand, it is questionable whether data protection law indeed has the sole or main purpose of protecting natural persons. Several authors point out that data protection law, at least initially, was mainly aimed at protecting objective legal principles and general interests. On the other hand, the question under discussion in the legal literature is precisely to what extent the protection of natural persons is the best basis for future regulation or should be extended to groups or society at large.

This research combines insights from the legal domain, from the technological domain and, in part, from societal developments.

The legal regime was assessed on three points.

(4) The current legal regime and the existing definitions, and explanations thereof in literature or authoritative opinions, were assessed to determine what the existing framework is for evaluating data processing.
(5) The history of the legal regime on the point of definitions was evaluated for three reasons. First, it shows how the data protection framework has been altered over time in response to societal and technological changes. Second, it gives insights in the logic and rationales behind the current definitions and categorisation: why are the definitions as they are, what do they aim to achieve? More in general, attention was paid to the discussion on the overarching rationale of the data protection framework, as such is relevant with an eye to potential future change made to the data protection framework. Third, through the various definitions and delineations of the data categories and especially, the variations discussed and contemplated in the legislative history yet rejected, alternative ways of approaching the regulation of data can be found.
(6) The potential future of data protection framework was assessed. The technological and societal developments discussed in this study have a considerable impact on the interpretation and effects of the current regulatory framework. That is why an overview is provided of the most important thoughts on the potential for altering the current regulatory framework.

The technological realm was assessed on three points:

(4) A brief overview of the technological developments after World War II was provided in order to paint the picture of a field constantly in flux. This description shows the background against which the legal framework was altered over time.
(5) It assessed the current technologies, especially in light of the various legal data categories and the boundaries between them. This description shows that it is increasingly possible to de-anonymise a dataset and to infer (sensitive) personal data from one or more aggregated datasets.
(6) It described technological developments that might change the landscape even further in the future. This shows that if anything, it the lines between the various legal data categories will be blurred to an even greater extent.

Also, attention was paid to two societal developments (though these are infused by both legal and technological developments):

(3) The study describes how technologies have become general available over time. This means that more and more governmental organisations, companies and even citizens have highly advanced technological resources at their disposal. The consequence of this trend is that if data

is shared between various parties or made publicly available, it is increasingly likely that there will be a party that will operate on it in a way that affects the legal status of the dataset.

(4) The study briefly points to the legal and societal push to make data publicly available. This concerns primarily concerns statistical data, public sector information and non-personal data. Mostly, these datasets will in and by themselves not contain personal data, but when combined with other datasets, they may be used to generate (sensitive) personal data. In addition, given the advancement and general availability of technologies, it is increasingly likely that there will be a party that will invest enough resources to deanonymize or reidentify a dataset.

1.6 Overview report

The overview of this report is as follows:

Chapter 1: Introduction
- Section 1.1 provides the introduction;
- Section 1.2 introduces the background of this study;
- Section 1.3 sets out the research questions that guide the study;
- Section 1.4 discusses the core methodologies used for this study;
- Section 1.5 provides the main themes and topics discussed in this report;
- Section 1.6 gives an overview of the report.

Chapter 2: Anonymization, de-anonymization, and non-personal data
- Section 2.1 gives the introduction;
- Section 2.2 discusses the legal distinction between anonymous and non-anonymous data;
- Section 2.3 describes the main techniques available for anonymising and de-anonymising data;
- Section 2.4 provides an analysis of the main conclusions of the chapter.

Chapter 3: Aggregation and composition
- Section 3.1 gives the introduction;
- Section 3.2 discusses the legal distinction between aggregated and non-aggregated data;
- Section 3.3 describes the main techniques available for aggregating and re-identifying data;
- Section 3.4 provides an analysis of the main conclusions of the chapter.

Chapter 4: Pseudonymization and de-pseudonymization
- Section 4.1 gives the introduction;
- Section 4.2 discusses the legal distinction between pseudonymous and non-pseudonymous data;
- Section 4.3 describes the main techniques available for pseudonymising and de-pseudonymising data;
- Section 4.4 provides an analysis of the main conclusions of the chapter.

Chapter 5: Sensitive and non-sensitive personal data
- Section 5.1 gives the introduction;
- Section 5.2 discusses the legal distinction between sensitive and non-sensitive personal data;
- Section 5.3 describes the main techniques available for inferring sensitive data form non-sensitive personal data and non-personal data;
- Section 5.4 provides an analysis of the main conclusions of the chapter.

Chapter 6: Analysis
- Section 6.1 gives the introduction;
- Section 6.2 provides the main conclusions of this study;
- Section 6.3 discusses the regulatory objective of privacy and data protection law;
- Section 6.4 discusses several ways to regulate data and the various trade-offs;
- Section 6.5 provides regulatory alternatives as found in academic literature;
- Section 6.6 sketches potential paths forward, divided over five ideal type scenarios;
- Section 6.7 answers the research questions for this study.

Chapter 7: Annexes
- Section 7.1 contains the full interview reports;
- Section 7.2 contains the full workshop reports;
- Section 7.3 contains an overview of the most relevant provisions in the GDPR.

The setup of chapters 2 to 5 is the same. Section 2 of each chapter contains a description of the EU and CoE regulatory approach. Section 3 of each chapter contains a literature overview from a technological perspective (subsection 3.1), the relevant results from the interviews (subsection 3.2) and the relevant results from the workshop (subsection 3.3). Section 4 provides an analysis by giving a summary of the main challenges and tension between the legal and the technical reality.

Finally, a caveat applies. This report is based on classic distinctions between categories of data as prevalent in Europe's privacy and data protection legislation. These distinctions guide the delineation between chapters 2-5. However, the precise reason for and background of this study is the fact that these distinctions are increasingly difficult to draw. Thus the findings from the literature study, interviews and workshop often have relevance for multiple chapters and multiple legal distinctions. The same applies, though to a lesser extent, to legal findings. For example, the EU's regulation on non-personal data is relevant for both chapters 3 and 4. Because non-European countries do not use the same legal distinctions as the EU, or draw the boundary between the different categories somewhere else than is common in EU legislation, findings are not always easily placed in one or the other chapter. In order to avoid duplications, these findings are only discussed once in this report, in the chapter for which the findings are most relevant or, when the findings are equally relevant for more than one chapter, in the first chapter for which it is relevant.

Chapter 2: Anonymization, de-anonymization and non-personal data

2.1 Introduction

This chapter describes the boundary between personal data and anonymised data. While through aggregation, data can be anonymised by treating the data no longer at the level of $n = 1$, but on the level of $n = 20$, $n = 100$, etc., this chapter concerns the possibility of stripping the data, still concerning one individual or a small group, of identifiable data. Instead of saying 'Bart van der Sloot lives in Amsterdam and owns a red Ferrari' one says 'person X lives in Amsterdam and owns a purple Ferrari'. Probably, this will be enough for the sentence to be considered non-personal data, but what if Bart van der Sloot lives in a city that does not have a million inhabitants like Amsterdam, but in one with 3.000 inhabitants, like the city of Rheden? In that case, there will most likely only be one person with a purple Ferrari, and so the sentence 'person X owning a purple Ferrari living in Rheden' may still be considered personal data because it allows others to identify a person based on the information, if only the other inhabitants of the town.

This chapter concerns both the process of anonymisation and de-anonymization and non-personal data, which may be both anonymised data and data that were never personal data. The next chapter will deal with the question of aggregation. Both anonymisation and aggregation can lead to the same result, the data are no longer personal data, but both are done through different techniques and have different vulnerabilities, which is why they are discussed separately in this study. Section 2.2 discusses the legal distinction between anonymous and non-anonymous data, section 2.3 describes the main techniques available for anonymising and de-anonymising data and section 2.4 analyses the gap between the legal regulation and technical reality.

2.2 Legal regulation

This section will briefly delve into the legislative history of the DPD (section 2.2.2) and the GDPR (section 2.2.5), and the legal interpretation of that regulation by the CJEU (section 2.2.1.3) and the Working Party 29 (section 2.2.4). It will also touch upon the Council of Europe's Convention 108+ (section 2.2.6) and the EU's Regulation on the free flow of non-personal data (section 2.2.7). Before these specific elements are discussed, it is important to sketch some of the contours of the right to data protection, as opposed to the right to privacy (section 2.2.1).

2.2.1 Data Protection Directive

The origins of the right to data protection lie partially in the data protection rules of northern European countries,[10] which arose in several nations in the 1970s, and the subsequent Resolutions on data processing by the Council of Europe[11] and partially in the USA and the realization of so-called Fair Information Principles. The increased use of large databases raised a number of problems for the traditional concept of the right to privacy. Perhaps most important, data processing often does not concern private or sensitive data, but public and non-sensitive data such as car ownership, postal codes, number of children, etc.[12] Such data processing are traditionally not considered to fall under the scope

[10] This section partly based on: Van der Sloot, B. (2014). Do data protection rules protect the individual and should they? An assessment of the proposed General Data Protection Regulation. International Data Privacy Law, 4(4), 307.
[11] Dammann, U. Mallmann, O. & Simitis, S. (eds) (1977)Data Protection Legislation: An International Documentation: Engl.–German: eine internationale Dokumentation = Die Gesetzgebung zum Datenschutz, Frankfurt am Main: Metzner. Hondius, F.W. (1975) Emerging Data Protection in Europe, Amsterdam: North-Holland.. Burkert, H. (1983). Freedom of Information and Data Protection, Bonn: Gesellschaft für Mathematik und Datenverarbeitung.
[12] Secretary's Advisory Committee on Automated Personal Data Systems (1973) Records, Computers and the Rights of Citizens.

of the right to privacy, as contained, inter alia, in the European Convention on Human Rights (ECHR) 1950, adopted by the Council of Europe. The ECHR, as well as the United Nations Universal Declaration of Human Rights 1948, were adopted in the wake of the Second World war. These human rights frameworks contain a right to privacy, as well as several other rights, and were intended primarily for vertical relations (protecting citizens from intrusive governments). They were intended to protect the citizen against highly intrusive forms of governmental interferences, such as body cavity searches, placing children out of home, and subjecting dissidents to permanent forms of surveillance. This means that processing of ordinary personal data and processing of personal data by other parties than governmental organisations cannot, or only indirectly, be addressed under the Convention and that the processing of data by governmental organisations will only fall under the scope of Article 8 ECHR if it either significantly affects a person's private life or can be considered an interference with her right to correspondence.

Article 8 - Right to respect for private and family life

1. Everyone has the right to respect for his private and family life, his home and his correspondence.
2. There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.

To address the new challenges of data processing, the CoE adopted two Resolutions for data processing in 1973 and 1974, one for the public and one for the private sector, which defined 'personal information' simply as information relating to individuals (physical persons).[13] This consequently meant a significant extension of the scope of protection offered to citizens, as the right to privacy, in principle, only related to the processing of sensitive personal data or to data processing that had a substantial impact on her private life or right to correspondence. Now, any processing of any data is regulated, as long as the data relates to a natural person. In 1981, the Council adopted the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, also known as Convention 108, in which 'personal data' were defined in a slightly broader fashion, namely as any information relating to an identified or identifiable individual.[14] In the Data Protection Directive of the European Union (EU) of 1995,[15] the material scope was widened even further. It not only introduces a very wide, non-exhaustive list of possible identifying factors, but also the possibility of 'indirect' identifiable data.[16] Finally, in the General Data Protection Regulation 2016, replacing the Data Protection Directive, personal data are defined in a slightly broader manner,[17] and in Convention 108+, the updated version of CoE's Covention form 1981, the definition of 'personal data' has remained the same.

Though, over time, the ECtHR has expanded the scope of the right to privacy in order to include many modern-day data processing operations, the material scope of the right to privacy (Article 8 ECHR) is still different from that of data protection law. The data protection regime has a wider scope of application for at least two reasons. First, the material scope is dependent on the definition of ´personal data´, which, as will become clear in this report, is particularly wide; though the term 'private life', contained in Article 8 ECHR is also wide, the scope of the two notions do not always overlap. That is: not all processing of personal data will be considered to affect a person's 'private life'. Second, in the human rights framework, a claim is assessed on both the *ratione materiae* (does the matter complained

[13] Council of Europe, Committee of Ministers, Resolution (73) 22 On the Protection of the privacy of individuals vis-à-vis electronic data banks in the private sector. (Adopted by the Committee of Ministers on 26 September 1973 at the 224th meeting of the Ministers' Deputies). Council of Europe. Committee of Ministers, Resolution (74) 29 On the Protection of the privacy of individuals vis-à-vis electronic data banks in the public sector. (Adopted by the Committee of Ministers on 20 September 1974 at the 236th meeting of the Ministers' Deputies).
[14] Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, Strasbourg, 28 January 1981, article 2 sub a.
[15] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.
[16] This was even broadened further: ECLI:EU:C:2013:355.
[17] GDPR, Article 4(1)..

of fall under the material scope of the article invoked?) and the *ratione personae* principle (can the applicant claim to be a victim?). With respect to that second question, there is a significant threshold, as applicants must be able to show that they have suffered from direct, individualizable, and substantial harm. Under the data protection framework, both principles are merged. This means that any processing of personal data, however mundane and small, even writing in a blog post 'Boris Johnson has blue eyes', is considered personal data processing, to which the GDPR applies.

On the other hand, the right to privacy, as contained in Article 8 ECHR, is broader than the right to data protection because its application does not depend on the question of whether personal data are processed. As will become clear in this report, many modern data processing operations operate at the border of the data protection framework and revolve around processing aggregated and group data, which may or may not fall under the scope of the GDPR. If, however, such processing affects the right to private life, which the ECtHR has interpreted in a fairly broad fashion, it will fall under the right to privacy. Importantly, the ECtHR has found that concepts such as personal autonomy, human dignity, and individual freedom underpin the right to privacy,[18] so that when processing initiatives affect those interests, the Court may find an interference with the right to privacy.[19]

Finally, it is important to note a difference between EU and CoE instruments in this field. Convention 108 explicitly allows the Member States to specify in their national legislation 'that it will also apply this Convention to information relating to groups of persons, associations, foundations, companies, corporations and any other bodies consisting directly or indirectly of individuals, whether or not such bodies possess legal personality';[20] the updated Convention, named Convention 108+, adopted in 2018, reemphasizes that principle: 'While the Convention concerns data processing relating to individuals, the Parties may extend the protection in their domestic law to data relating to legal persons in order to protect their legitimate interests. The Convention applies to living individuals: it is not meant to apply to personal data relating to deceased persons. However, this does not prevent Parties from extending the protection to deceased persons.'[21] The EU's legislative frameworks are narrower in this sense: in principle, they only apply to personal data about natural (living) persons.[22]

For non-Europeans, it is important to understand that there are two supranational organizations in Europe. The Council of Europe (Convention 108+ and ECHR) and the European Union (GDPR and Charter of Fundamental Rights). The European Court of Justice (CJEU) is the highest court of the European Union. Twenty-seven countries are members of the European Union. The CoE's ECHR is overseen by the ECtHR. Forty-seven of the about fifty European countries are members of the Council of Europe. All EU Member States are also members of the Council of Europe. EU laws, such as the GDPR, take precedence over national laws. If, for example, if Italian law conflicts with the GDPR, the Italian law would be declared invalid, and the GDPR would take precedence. The same counts for the rulings of the CJEU and the ECtHR: they take precedence over national laws and courts.

Initially, the division of tasks between the Council of Europe and the European Union was clear: the Council of Europe focused on protecting human rights, while the EU, as the successor of the *European Coal and Steel Community*, was mainly concerned with economic and socio-economic issues. Gradually, however, the European Union has adopted rules and regulations on almost every aspect of society, including human rights. The Charter of Fundamental Rights of the European Union from 2000 entered into force in 2009[23] and can be seen as the constitution of the European Union. Like the

[18] Van der Sloot, B. (2014). Privacy as human flourishing: Could a shift towards virtue ethics strengthen privacy protection in the age of Big Data. *J. Intell. Prop. Info. Tech. & Elec. Com. L.*, *5*, 230.
[19] An interference is not a violation. Like when personal data are processed, this means that the GDPR applies, but this does not mean that data processing is prohibited. Similarly, an interference with a human right can be legitimate, when such is deemed necessary in a democratic society and had a basis in a law.
[20] Convention 108, 1981, Article 3.2(c).
[21] <https://rm.coe.int/convention-108-convention-for-the-protection-of-individuals-with-regar/16808b36f1>.
[22] GDPR, Article 4(1).
[23] See also: Van der Sloot, B (2014). 'Do data protection rules protect the individual and should they? An assessment of the proposed General Data

European Convention on Human Rights, it contains rights such as freedom of religion, freedom of speech, and the right to privacy. Importantly, it adds a number of rights to the catalogue of rights contained in the ECHR, such as the right to data protection. The Charter contains provision for the right to data protection (Article 8) which is separated from the right to privacy (Article 7).

> Article 7 - Respect for private and family life
>
> Everyone has the right to respect for his or her private and family life, home and communications.
>
> Article 8 - Protection of personal data
>
> 1. Everyone has the right to the protection of personal data concerning him or her.
> 2. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified.
> 3. Compliance with these rules shall be subject to control by an independent authority.

Although data protection instruments were introduced to complement the right to privacy, early data protection instruments were explicitly linked to the right to privacy, and the right to data protection was seen either as a sub-set of privacy interests or as a twin right. The two resolutions by Council of Europe from 1973 and 1974 were titled Resolution 'on the protection of the privacy of individuals vis-à-vis electronic data banks in the private sector' and Resolution 'on the protection of the privacy of individuals vis-à-vis electronic data banks in the public sector'. Data processing issues were still explicitly seen as a part of or following from the right to privacy. The Resolution on the public sector also stated explicitly 'that the use of electronic data banks by public authorities has given rise to increasing concern about the protection of the privacy of individuals'. Convention 108 does not contain the word privacy in its title but specified in its preamble: 'Considering that it is desirable to extend the safeguards for everyone's rights and fundamental freedoms, and in particular the right to the respect for privacy, taking account of the increasing flow across frontiers of personal data undergoing automatic processing; Reaffirming at the same time their commitment to freedom of information regardless of frontiers; Recognising that it is necessary to reconcile the fundamental values of the respect for privacy and the free flow of information between peoples'. Also, Article 1 of the Convention, laying down the object and purpose of the instrument, made explicit reference to the right to privacy: 'The purpose of this Convention is to secure in the territory of each Party [each member state to the Council of Europe] for every individual, whatever his nationality or residence, respect for his rights and fundamental freedoms, and in particular his right to privacy, with regard to automatic processing of personal data relating to him ("data protection").' The explanatory memorandum to the Convention mentioned the right to privacy a dozen times. Thus, although the reference to privacy in the title was omitted, it is still obvious that the rules on data protection, as laid down in the Convention, must be seen in light of the right to privacy. Also, under the European Convention on Human Rights, data protection issues are still seen as a subset of privacy, similar to most parts of the world, where these are considered to fall under the notion of 'informational privacy'.

This is different in the EU, where the right to privacy and the right to data protection are considered independent fundamental rights. The EU started to engage in the field of data protection in the late eighties and early nineties of the previous century. The original mandate to regulate data protection by the EU was also found in market regulation. The GDPR, which aims at protecting the fundamental right to data protection as specified in Article 8 of the Charter, is still only partly an instrument protecting a human right, and certainly also, and arguably predominantly, an instrument regulating the data market.

Protection Regulation', International Data Privacy Law, 3.

Still, even in the EU, the right to data protection was initially strongly connected to the right to privacy,[24] which was reflected in the Directive, which makes reference to the right to privacy 13 times and in Article 1, concerning the objective of the Directive, holds: 'In accordance with this Directive, Member States shall protect the fundamental rights and freedoms of natural persons, and in particular their right to privacy with respect to the processing of personal data. Member States shall neither restrict nor prohibit the free flow of personal data between Member States for reasons connected with the protection afforded under paragraph 1.' However, in the General Data Protection Regulation, the reference to privacy has been deleted entirely. Established terms such as 'privacy by design' have been renamed to 'data protection by design', and 'privacy impact assessments' have become 'data protection impact assessments'. The objective of the Regulation, in Article 1, refers to the protection of personal data and not of privacy.

2.2.2 Data Protection Directive

In the Commission's proposal for the Directive, the notion of personal data was redefined, and the notion of depersonalisation was added.

Depersonalisation is, in a certain sense, the predecessor of the notion 'anonymisation'. Depersonalisation, perhaps a clearer and more accurate term than the latter, was interpreted as meaning modifying information in such a way that it could no longer be associated with a specific individual. The definition contained a contextual element, as emphasized by the explanatory memorandum: 'An item of data can be regarded as depersonalized even if it could theoretically be repersonalized with the help of disproportionate technical and financial resources' and by the definition of depersonalization, which 'means modify personal data in such a way that the information they contain can no longer be associated with a specific individual or an individual capable of being determined except at the price of an excessive effort.'[25]

The definition of personal data was extended so as to include 'indirect' identifiable data. 'In order to avoid a situation in which means of Indirect Identification make it possible to circumvent this definition, it is stated that an Identifiable Individual is an Individual who can be identified by reference to a number or a similar Identifying particular.'

The Economic and Social Committee stressed that the definition of depersonalisation was clearer than the explanation given in the memorandum. 'The explanation limits the scope of the definition, allowing further attention to be given to data which, although depersonalized by their producer, remain associated, after communication, with personal data from other processing. Moreover, 'excessive effort' should be deleted, for a processing task requiring an excessive effort today may require no effort at all next year.'[26] The latter remark is interesting because it is still valid today. It would also remove part of the vagueness, fluidity, and contextuality of the definition of depersonalization or, mutatis mutandis, of anonymisation. Accordingly, parliament suggested a new definition: ''depersonalize' means modify personal data in such a way that the information they contain can no longer be associated with a specific individual or an individual capable of being determined'.[27] Depersonalisation, according to parliament, also included aggregation, as was evidenced by its proposal to include in the Directive a rule that a data controller could disclose data 'for research and statistical purposes on condition that the personal data is depersonalized.'

The Commission then came up with an amended proposal in which it deleted the concept of depersonalisation. It did not replace that with a general reference to anonymisation but did include in

[24] Gonzalez Fuster, G (2014). The emergence of personal data protection as a fundamental right of the EU, Springer: Dordrecht..
[25] COM(90) 314 final ~.sYN 287 and 288 Brussels, 13 September 1990.
[26] C 159 Volume 34 17 June 1991.
[27] No C 94/ 176 Wednesday, 11 March 1992.

the definition of 'personal data' a reference to aggregated data, thus including part of the anonymised data, but not all. The definition it proposed was: ''personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity; data presented in statistical form, which is of such a type that the persons concerned can no longer be reasonably identified, are not considered as personal data'.[28] Later on, the reference to statistical data was excluded from the definition.

The Council was the first to introduce the concept of 'anonymity' in a consideration. 'Whereas the principles of protection must apply to any information concerning an identified or identifiable person; whereas, to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person; whereas the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable; whereas codes of conduct within the meaning of Article 27 may be a useful instrument in providing guidance as to the way in which data may be rendered anonymous and retained in a form in which identification of the data subject is no longer possible'. This is similar to the recital adopted in the final version of the DPD: 'Whereas the principles of protection must apply to any information concerning an identified or identifiable person; whereas, to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person; whereas the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable; whereas codes of conduct within the meaning of Article 27 may be a useful instrument for providing guidance as to the ways in which data may be rendered anonymous and retained in a form in which identification of the data subject is no longer possible'.[29]

A clear explanation of these subsequent decisions is lacking, but what appears to be clear is that the definition and especially the explanation of the definition of 'depersonalisation' was considered too vague and too abstract, that in response it was made more objective, replaced by a reference to aggregated data and then finally deleted. But the same type of formulation later reappears, though not in the operative part of the Directive but in one of its recitals under the term anonymization. Perhaps if anything, the terminology used is more contextual, as the initial definition of depersonalization contained a reference to 'at the price of an excessive effort', while the recital that was adopted refers to 'all the means likely reasonably to be used either by the controller or by any other person'. The reference to all means likely seems more open than excessive effort; in addition, it explicitly recognizes that account should not only be taken of the means likely used by the initial data controller, but by anyone. What must be concluded is that a more objective and non-contextual definition of depersonalization (the revised definition) was put on the table, but was rejected in favour of a more broad, fluid, and contextual formulation.

2.2.3 CJEU

The CJEU has issued a number of judgements relevant to the definition of personal data in general and anonymisation in particular. A selection of the most important decisions will be highlighted briefly below. Importantly, many judgements of the CJEU with respect to anonymisation concern the e-Privacy regime, which falls outside the scope of this study.[30] Three topics that were discussed in the

[28] No C 311 / 38 Official Journal of the European Communities 27 November 1992.
[29] DPD, Recital 26..
[30] See e.g. CJEU, C-275/06, *Productores de Música de España (Promusicae) v Telefónica de España SAU* [2008] ECLI:EU:C:2008:54; CJEU, C-461/10, *Bonnier Audio AB and Others v Perfect Communication Sweden AB* [2012] ECLI:EU:C:2012:219; CJEU, C-203/15,*Tele2 Sverige AB v Post- och telestyrelsen and Secretary of State for the Home Department v Tom Watson and Others* [2016] ECLI:EU:C:2016:970; and CJEU, C-398/15, *Camera di Commercio, Industria, Artigianato e Agricoltura di Lecce v Salvatore Manni* [2017] ECLI:EU:C:2017:197; CJEU, C-623/17,*Privacy International v Secretary of State for Foreign and Commonwealth Affairs and Others* [2020] ECLI:EU:C:2020:790; CJEU, C-511/18, *La Quadrature du Net and Others*

jurisprudence will be discussed below: the publication of non-anonymised data, the identifiability of cookies and IP addresses, and the rights of data subjects vis-à-vis anonymised data.

2.2.3.1 Publication of non-anonymised data

In *Österreichischer Rundfunk*, there was discussion over the need to anonymize documents before disclosing them. The Court stressed that it is allowed to publish names and intimate details of persons when proportionate and necessary. It found that the DPD does not preclude national legislation requiring such publication, provided that it is shown that 'the wide disclosure not merely of the amounts of the annual income above a certain threshold of persons employed by the bodies subject to control by the Rechnungshof but also of the names of the recipients of that income is necessary for and appropriate to the objective of proper management of public funds pursued by the legislature, that being for the national courts to ascertain.'[31]

In the *Borax* case, the Commission did not want to disclose expert opinions it had received because this would jeopardize their privacy and undermine the sphere of confidentiality. Again, however, the CJEU ruled differently, stressing that scientific opinions obtained by an institution for the purpose of the preparation of legislation must, as a rule, be disclosed. This is so, the CJEU found, even if they might give rise to controversy or deter those who expressed them from making their contribution to the decision-making process of that institution. It recognized that the risk that public debate born of the disclosure of their opinions may deter experts from taking further part in its decision-making process is inherent in that rule. However, that risk cannot lead to the principle that any disclosure of a scientific opinion with significant consequences will have a deterrent effect as regards its author or that the risk is such as seriously to undermine the institution's decision-making process, as where that institution would find it impossible to consult other experts.[32]

Thus, the CJEU makes clear that there is no obligation to anonymize data, even when making them public. The publication of personal data may be legitimate when there is a legitimate processing ground (in both cases there was a legal obligation) and such is proportionate and necessary in light of a legitimate aim. Admittedly, both cases did not concern the publication of sensitive personal data, which may have led to a different ruling by the CJEU..

2.2.3.2 IP-addresses

The case of *Scarlet* concerned the question of whether IP addresses should be considered personal data in and by themselves. The CJEU confirmed that they could, but depending on the the circumstances of the case. Thus, not all IP addresses can be considered personal data. For example, an IP address may be personal data when it allows a person to be identified by reference to an identification number or any other information. It thus pointed out that it is not so much relevant to determine the legal status of IP addresses as it is to determine the circumstances in which and the purposes for which they may be collected, the circumstances in which the resulting personal data may be resolved and processed, or even the conditions under which their collection and resolution may be requested. It underlined that a filtering and blocking system was, without question, likely to affect the right to protection of personal data to a sufficient degree to enable it to be classified as a limitation within the meaning of Article 52(1) of the Charter.[33] Thus, it adopted a contextual approach to determining whether data are personal data. In the *Breyer* case, the Court reaffirmed that position with respect to dynamic IP addresses.[34]

---

v Premier ministre and Others [2020] ECLI:EU:C:2020:791 and ECLI:EU:C:2021:152; and CJEU, C-597/19, *Mircom International Content Management & Consulting (M. I. C. M.) Limited contra Telenet BVBA* [2021] ECLI:EU:C:2021:492.

[31] CJEU, C-465/00, *Österreichischer Rundfunk and Others* [2003] ECLI:EU:C:2003:294, para. 94.

[32] CJEU, T-121/05, *Borax Europe Ltd v Commission of the European Communities* [2009] ECLI:EU:T:2009:64. See also later: CJEU, T-483/13, *Athanassios Oikonomopoulos v European Commission* [2016] ECLI:EU:T:2016:421.

[33] CJEU, C-70/10, *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)* [2011] ECLI:EU:C:2011:255.

[34] CJEU, Case C-582/14, *Patrick Breyer v Bundesrepublik Deutschlandv* [2016] ECLI:EU:C:2016:779.

2.2.3.3 Data subject rights

If an organization has anonymised personal data, data subjects can no longer invoke their rights, not only because, juridically speaking, the data protection framework no longer applies, but also, practically speaking, because data can no longer be linked data to the data subject invoking its rights (e.g. right to access). Does that mean that data should not be anonymized?

As specified in *Rijkeboer* judgement this may not be the case.[35] In this judgement, the CJEU required, again, a contextual analysis of all interests at stake which, in accordance with the circumstances of the particular case, those were equivalent to the interests of the data controller and the data subject. The Court established that 'Article 12(a) of Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data requires Member States to ensure a right of access to information on the recipients or categories of recipient of personal data and on the content of the data disclosed not only in respect of the present but also in respect of the past. It is for Member States to fix a time-limit for storage of that information and to provide for access to that information which constitutes a fair balance between, on the one hand, the interest of the data subject in protecting his privacy, in particular by way of his rights to object and to bring legal proceedings and, on the other, the burden which the obligation to store that information represents for the controller. Rules limiting the storage of information on the recipients or categories of recipient of personal data and on the content of the data disclosed to a period of one year and correspondingly limiting access to that information, while basic data is stored for a much longer period, do not constitute a fair balance of the interest and obligation at issue, unless it can be shown that longer storage of that information would constitute an excessive burden on the controller. It is, however, for national courts to make the determinations necessary.'[36] This principle was later formalized in Article 11 GDPR.

From another perspective, in the *OC* judgement, the Court analysed the data protection concerns of a press release were the identity of the individual had been pseudonymised.[37] As extracted from the case, reidentification of the individual was possible by means of other data available online. This led the complainant to claim that a reader, using the details of the press release, would be able to reidentify the complainant without any difficulty, since, among the 70 projects listed on the press release, only three involved funding of around 1.1 million euros and were run by women. Based on these data, an internet search allowed the complainant to be reidentified among the other women, in particular because her father worked at the university hosting the project. In a remarkable passage, the Court found that the organisation cannot be held accountable for other information put online and could thus only be held accountable for its own press release. Although the Court acknowledged that the press release contained the gender, the approximate age, occupation, and nationality of the complainant, and although it referred to the father of the complainant in question and to the place where his profession was exercised, the approximate amount of the grant, the granting body, the nature of the entity hosting the project and its geographical location, the Court, notwithstanding, did not find that the press release contained personal data. Important for this outcome was the fact that it would have been necessary to go through the description of each of the 70 projects listed on the website to understand who was the scientist in charge of the project, the host institution, and the amount of funding, which would necessarily have taken considerable time and would not have been so easy. Even though journalists could identify the complainantr, the Court established that such result could not be attributed to OLAF, the organisation in charge of issuing the press statement. As out forward by the Court,'[i]t follows that the German

[35] CJEU, C-553/07, *College van burgemeester en wethouders van Rotterdam v M. E. E. Rijkeboer* [2009] ECLI:EU:C:2009:293
[36] CJEU, *Rijkeboer,* para. 70.
[37] CJEU, T-384/20, *OC v European Commission* [2022] ECLI:EU:T:2022:273.

journalist and the Greek journalist were not able to identify the applicant based solely on the identifiers present in press release No. 13/2020 and that, in any event, it was necessary for them, in one way or another, to use external and complementary identification elements to said press release. In this regard, it is important to recall the case-law, cited in paragraph 49 above, according to which only acts or conduct attributable to an institution or body of the Union may give rise to the liability of the Union, so that the elements of information taken outside of the said press release cannot serve as a basis for the engagement of its responsibility'.[38]

2.2.4 Article 29 Working Party

Article 29 Working Party has issued many relevant opinions on the scope of personal data and anonymization. Here, the opinions most directly related to these topics will be briefly discussed, namely the Opinion 4/2007 on the concept of personal data and three other opinions that touch upon anonymity.

2.2.4.1 Personal data

Article 29 Working Party issued an opinion on the concept of personal data.[39] To determine whether data should be considered personal, four factors should be taken into account, each of which should be interpreted broadly:

1. Any information:
   - Nature: 'personal data' may be objective information and subjective information, qualifications, and expectations (e.g. Bob is probably going to die soon). In addition, it may concern incorrect data (e.g. Boris Johnson is the leader of the Labour Party), when the individual can be clearly identified.
   - Content: 'personal data' may refer to information concerning who a person is, what she does, how she feels, what her capacities are, and any other types of information about a specific person.
   - Format: the medium or carrier of the information is not important for the definition of personal data. Thus, it may refer to information in any form, such as alphabetical, numerical, graphical, photographical, or acoustic. Most importantly, the data protection framework applies to personal data processed through automated means or through non-automated means when such data are structured (e.g. a paper archive).[40]
2. Relating to:
   - Content: information relates to a person when the content refers to or relates to a person. For instance, if the results of a medical analysis establish that Bob is ill, said information obviously relates to Bob in content.
   - Purpose: information can also relate to a person depending on its purpose. For example, because Bob is a man between 20-35, the information relating to the age of Bob can be used by his car insurance company to establish the corresponding insurance fee.
   - Result: '[a] third kind of 'relating' to specific persons arises when a "result" element is present. Despite the absence of a "content" or "purpose" element, data can be considered to "relate" to an individual because their use is likely to have an impact on a certain person's rights and interests, taking into account all the circumstances surrounding the precise case. It should be noted that it is not necessary that the potential result be a major impact. It is sufficient if the individual may be treated differently from other persons as a result of the processing of such data.'[41]
3. Identified or identifiable:

[38] CJEU, *OC v European Commission,* para 87.
[39] Article 29 Working Party, 'Opinion 4/2007 on the concept of personal data', 01248/07/EN, 20 June 2007.
[40] GDPR, Article 2 para 1.
[41] Article 29 Working Party, 'Opinion 4/2007 on the concept of personal data', 11.

- Direct and indirect: Personal data includes both direct and indirect identifiable data.
- Identified and identifiable: Personal data includes both data through which a person can be identified and through which it is identifiable.
- Singling out: Importantly, it is not required that a data controller knows the name or identity of a person, as long as said person can be singled out.
- All the means likely reasonably: The Working Party emphasizes that the costs are a relevant factor, but not the only one. Time and the development of future re-identification and de-anonymization techniques are certainly important factors as well. 'Identification may not be possible today with all the means likely reasonably to be used today. If the data are intended to be stored for one month, identification may not be anticipated to be possible during the "lifetime" of the information, and said data should not be considered personal data. However, if the data are intended to be kept for 10 years, the controller should consider the possibility of identification that may occur also in the ninth year of their lifetime, and which may make them personal data at that moment. Controllers should be able to adapt to these developments as they happen, and incorporate appropriate technical and organisational measures in due course.'

4. Natural person:
   - Dead: Information about deceased persons, in principle, falls outside the scope of the data protection framework, but not if those data also say something about living persons.
   - Unborn children: In principle, the data protection framework does not apply to data about unborn children, but there is a complex discussion over what should be considered a living human being in this respect. In addition, most likely, the data collected about unborn children will, when stored, become personal data when the baby is born.
   - Legal persons: Information about legal persons will only be considered personal data if it indirectly says something about natural persons, such as the employees, owners, or shareholders. The boundaries between what should and what should not be considered personal data, again, are contextual..

2.2.4.2 Anonymisation

The Working Party has also issued a number of opinions on anonymisation.

The first opinion on anonymity on the internet dates from 1998. Article 29 Working Party primarily emphasised the importance of being able to browse the internet anonymously, although it agreed that there should be limits in certain environments.[42]

In 2000, in an opinion on Privacy on the Internet, the Working Party reemphasized that point and also discussed a number of technologies that could aid anonymous browsing.[43] Interestingly, it devoted little attention to the dangers of re-identification or de-anonymisation.

Finally, in 2014, the Working Party issued a new opinion on anonymisation techniques.[44] It argued that each technique should be evaluated on the basis of three criteria:

- is it still possible to single out an individual?;
- is it still possible to link records relating to an individual?; and
- can information be inferred concerning an individual?

It again found that the interpretation of 'anonymisation' under the data protection framework is necessarily a contextual one: 'Importance should be attached to contextual elements: account must be

[42] Article 29 Working Party, 'Anonymity on the Internet', 03 December 1997.
[43] Article 29 Working Party, 'Working Document Privacy on the Internet - An integrated EU Approach to On-line Data Protection', 21 November 2000.
[44] Article 29 Working Party, 'Anonymisation Techniques', 10 April 2014.

taken of "all" the means "likely reasonably" to be used for identification by the controller and third parties, paying special attention to what has lately become, in the current state of technology, "likely reasonably" (given the increase in computational power and tools available)."[45]

It stressed two key factors, which are worthwhile discussing because they are interpreted very narrowly by the Working Party:

- Data controllers should focus on the concrete means that would be necessary to reverse the anonymisation technique, notably regarding the cost and the know-how needed to implement those means and the assessment of their likelihood and severity. For instance, data controllers should balance their anonymisation effort and costs (in terms of both time and resources required) against the increasing low-cost availability of technical means to identify individuals in datasets, the increasing public availability of other datasets (such as those made available in connection with 'Open data' policies), and the many examples of incomplete anonymisation entailing subsequent adverse, sometimes irreparable effects on data subjects. It noted that the identification risk may increase over time and also depends on the development of information and communication technology. Legal regulations must therefore be formulated in a technologically neutral manner and ideally take into account the changes in the developing potentials of information technology.
- Secondly, 'the means likely reasonably to be used to determine whether a person is identifiable' are those to be used 'by the controller or by any other person.' This means that only if the data controller would aggregate the data to a level where the individual events are no longer identifiable, the resulting dataset can be qualified as anonymous. For example: if an organisation collects data on individual travel movements, the individual travel patterns at event level would still qualify as personal data for any party, as long as the data controller (or any other party) still has access to the original raw data. But if the data controller would delete the raw data, and only provide aggregate statistics to third parties on a high level, such as 'on Mondays on trajectory X there are 160% more passengers than on Tuesdays', that would qualify as anonymous data.[46]

Finally, the Opinion discussed the merits and pitfalls of various anonymization techniques. Firstly, it made a difference between two types of techniques. Generalization techniques, namely aggregation, k-anonymity, l-diversity, and t-closeness, which will be discussed in the next chapter. Secondly, it discussed several ways to arrive at, what it called, randomization, which is understood as a family of techniques that alters the veracity of the data in order to remove the strong link between the data and the individual. It discussed three techniques for arriving at randomization. Interestingly, each of those are not strictly anonymisation techniques in and by themselves, the Working Party emphasizes.

- Noise addition: adding noise to a dataset or making the data less accurate can help disabling third parties from using the data to identify a person. But Article 29 Working Party also warned that even though noise is added, it may still be possible for a third party to single out a person, even though this is perhaps more difficult. In some cases, attributing wrong or inaccurate data to a data subject may even be more harmful than when such data is accurate. Also, it made clear that adding noise is not, in itself, enough to anonymize a dataset. Data should be removed as well.
- Permutation: this technique pertains to the shuffling of the values of attributes in a table so that some of them are artificially linked to different data subjects. Again, this technique is not enough to anonymize data in and by itself and again, the Working Party warns that attributing incorrect data to data subjects may be harmful.
- Differential privacy: differential privacy can be used when the data controller generates anonymized views of a dataset whilst retaining a copy of the original data. Thus, those views or subsets are anonymous, but the data controller often still holds identifying information. Such

[45] idem., 6.
[46] idem., 9.

anonymized views would typically be generated through a subset of queries for a particular third party.

2.2.5 GDPR

Under the GDPR, a highly intense debate took place when drafting the Regulation on its scope and the defining concepts. In the end, the definition of personal data was altered only marginally, making reference to more examples through which a person can be identified. The recital referring to anonymization is defined somewhat more broadly in the GDPR than under the DPD Recital 26 makes explicit reference to singling out a person as a way of identification; in addition, it refers to many of the factors also identified by the Working Party. Although relatively little has consequently changed, it is important to briefly highlight some suggestions that were proposed but then rejected, as they may serve as inspiration for the regulatory alternatives discussed in chapter 6.

- First, there was a suggestion to include a rule that data could only be considered 'anonymous' when it was 'demonstrable' that they could no longer be linked to a person, thus placing a burden of proof on the data controller.
- Second, there was a proposal to exclude from the scope of personal data about a person's identity when concerning her professional capacity.
- Third, there were proposals to limit the scope of personal data, for example, restricting it to information through which a data subject can be 'unequivocally identified, directly or indirectly, by means available to the controller'.
- Fourth, there were proposals to extend the data protection framework to data about deceased persons.
- Fifth, there was a proposal to forbid re-identification: 'Reidentification of personal data, for instance by using retained online traces for the creation of profiles of the individuals, breaches of pseudonym and identification of the data subjects should be forbidden.'

2.2.6 Convention 108+

In Convention 108+, much of the expanded scope of personal data in the EU legal context has been adopted, though not by changing the formal definition, but by referring to those aspects in the explanatory memorandum. For example, it accepted that when a data controller is capable of individualising or singling out a person through information, such will also qualify as personal data. It suggests that such individualisation may be done through IP addresses, devices and other identifiers. The CoE also seems to adopt a similar contextual approach to anonymisation, but it seems even more mindful of the dangers of re-identification. It stressed that data could be considered anonymous only as long as it is impossible to re-identify the data subject or if such re-identification would require unreasonable time, effort, or resources, taking into consideration the available technology at the time of the processing and technological developments. Where it is possible for the controller or any person to identify the individual through the combination of different types of data, such as physical, physiological, genetic, economic, or social data, such data qualifies as personal data. 'When data is made anonymous, appropriate means should be put in place to avoid re-identification of data subjects, in particular, all technical means should be implemented in order to guarantee that the individual is not, or is no longer, identifiable. They should be regularly re-evaluated in light of the fast pace of technological development.'[47]

2.2.7 EU Regulation on the free flow of non-personal data

[47] Convention 108+, Explanatory memorandum.

Finally, it is important to stress that the EU has reinforced its clear distinction between personal data and non-personal data by adopting the Regulation on the free flow of non-personal data. While the General Data Protection Regulation arguably sets the highest level of protection in the world for personal data, the Regulation on the free flow of non-personal data takes a diametrically different approach. Processing non-personal data should not be restricted and is not bound by rules or limitations, the Regulation holds. Rather, the goal of this regulation is to remove the limits set in place by both governments and private parties to restrict the free flow of non-personal data.[48] This has made the distinction between non-personal and personal data even more relevant.

2.3 Technical developments

This section will provide insights gained on the technologies that can be used for anonymising and de-anonymising gained through literature study (section 2.3.1), the interviews conducted for this study (section 2.3.2) and a workshop held for this study (section 2.3.3).

2.3.1 Literature study

This section will focus on the anonymisation techniques in micro datasets (section 2.3.1.1) and the privacy models for microdata releases (section 2.3.1.2).

2.3.1.1 Anonymization techniques in microdata releases

Microdata, on a very basic level, is data that is person-specific. As such, microdata translates to a narrow scope of the individual by providing precise information about him or her. Examples of microdata encompass a person's social network profile (containing the email, first name, last name, age, location, etc. of the user), a health record (containing the name, age, social security number, and disease of a patient), or a driving license record (containing the name, gender, the identification number, vehicle, etc. of a driver). Given the accumulation of various data types in microdata sets, especially those containing sensitive information about the individual, microdata releases play a prominent role in privacy and data protection discourse.

The first attempt to delineate anonymization in microdata releases was made in the seminal work of Dalenius.[49] Dalenius considered that access to the released microdata should not result in increased knowledge about a specific individual. Its theory relied hence on the assumption that prior and posterior information about an individual contained in a database should remain similar. Since the notion of personal data is based on the distinguishability of an individual, the assumption introduced by Dalenius may hold certain properties akin to the legal conceptualization of personal data. For instance, it may provide solutions for the confidentiality of the released microdata, as it aims to prevent the acquisition of further knowledge about an individual. However, it is not fully aligned with the definition of personal data as conceived from a data protection perspective. Since the legal definition of personal data puts its onus on the identification of the individual, *i.e.* its distinguishability within a given group, the premises of Dalenius fall short of explaining the conceptualization of personal data from a legal perspective.

The singling out of an individual may be, in certain cases, but certainly not in all cases, a precondition for the gaining of knowledge about an individual, as postulated by Dalenius. For instance, the identification of an individual can be conceived as a sufficient condition prior to the furthering of knowledge about said individual in light of its related indirect identifiers. However, it may not be a necessary condition for gaining certain knowledge about an individual ifspecific attributes pertaining to him or her can be inferred based on indirect identifiers, *e.g.* where indirect identifiers possess a low

[48] Graef, I., Gellert, R., & Husovec, M. (2018). Towards a holistic regulatory approach for the European data economy: Why the illusive notion of non-personal data is counterproductive to data innovation.
[49] Dalenius, T. (1977). 'Towards a methodology for statistical disclosure control', Statistik Tidskrift, 15:429–444.

variability that may lead to reasonable estimations about an individual, even if it is not possible to single him or her out. Based on this assumption, the approach of Dalenius to anonymization has been considered narrow. For instance, it has been argued that Dalenius perspective does not contemplate the presence of additional information that an intruder may have in addition to the released dataset.

The crux of the matter is hence whether anonymization, as conceived in the GDPR, should be based just on the *singling out* of an individual, or if it should also cover the knowledge gain about a non-identified individual based on the disclosure of attributes relating to said individual. For example, imagine the release of a dataset containing the salaries of certain professionals. Three main identifiers are available in the original dataset, namely, the social security number (direct identifier), the salary (indirect identifier), and the job title (indirect identifier). If the social security numbers (direct identifiers) are erased, and the salaries (indirect identifiers) are anonymized (through the use of generalization, for instance), a hypothetical intruder may still determine the lower and upper bands of a given individual as long as the individual is contained in the database and the intruder knows his job title. This information may be further used to, among others, target a group of individuals to which the desired target forms part. Attribute disclosure in this regard is an area where the European data protection framework lacks pronouncement, as the legal definition of personal is subordinated to distinguishability.

The risk of re-identification in data releases, as well as the potential concurrence of attacks, should be minimized and controlled in order to render personal data anonymous. Most of the technical developments addressed in the literature deal with these issues and propose anonymization techniques and privacy models to ensure the anonymization of personal data.
There are different techniques and models to anonymize these datasets.[50] For example, entities that release microdata should consider not publishing the original micro-dataset *X*, but a modified version *Y*. The dataset *Y* is called the anonymized version of *X.* The following Figure 1 is a non-exhaustive list of various anonymization techniques. In the following, the various anonymization techniques are presented and explained.

This discussion draws heavily from the work done by Domingo-Ferrer et. al. on anonymization techniques and privacy models on microdata releases.[51]

[50] Agencia Española de Protección de Datos. 10 Misunderstandings related to Anonymisation, p. 5, available at: <https://edps.europa.eu/system/files/2021-04/21-04-27_aepd-edps_anonymisation_en_5.pdf>.
[51] Domingo-Ferrer, J., Sánchez, D., & Soria-Comas, J. (2016). Database anonymization: privacy models, data utility, and microaggregation-based inter-model connections. Synthesis Lectures on Information Security, Privacy, & Trust, 8(1), 1-136.

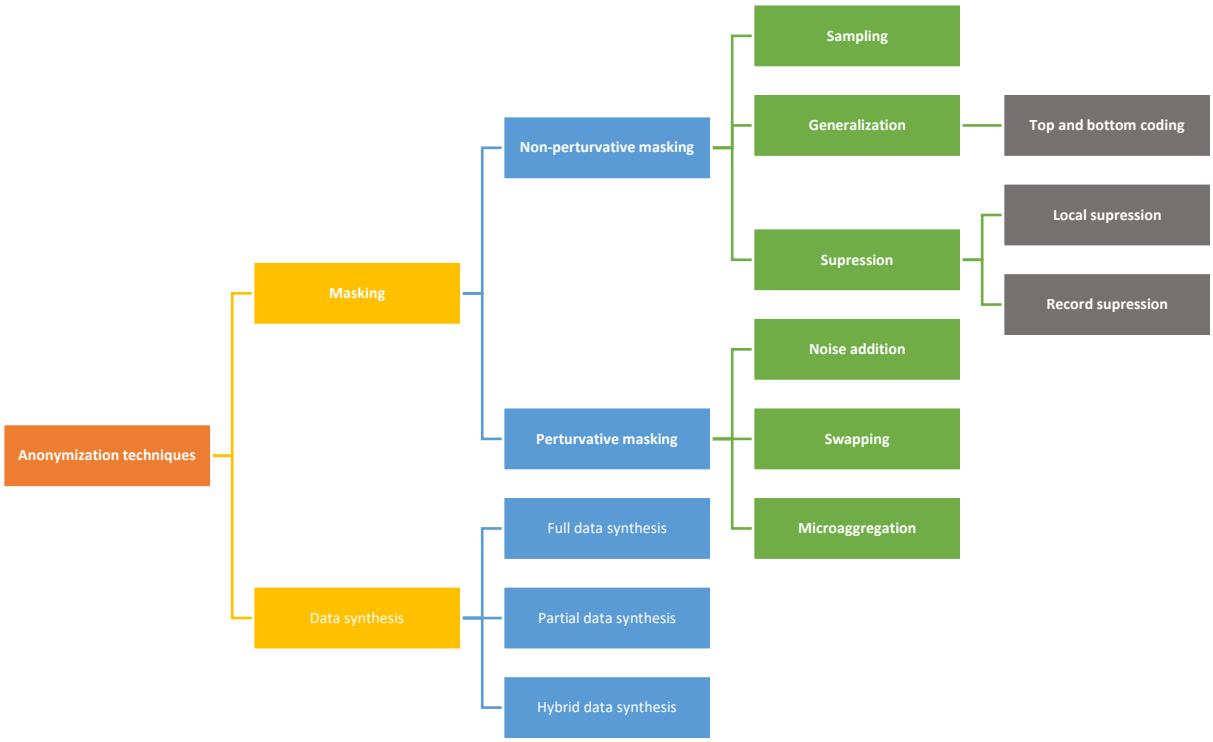

Figure 1: Anonymization techniques

Each of these techniques will be briefly discussed, the masking techniques (section 2.3.1.1.1) and then the synthetic data (section 2.3.1.1.2).

2.3.1.1.1 Masking techniques

Masking: this technique aims at inducing a relation between the original record *X* and the generated record *Y,* so that the indirect identifier is masked, which yields anonymity as a result.[52] Masking methods can be divided into two categories depending on their effects on the original record *X*.

- Non-perturbative masking: this technique focuses on the partial suppression or reduction of detail or coarsening of the original dataset *X*. As a result, dataset *Y* is not a perturbed dataset *per se,* but rather a reduced version of dataset *X,* which yields anonymity as a result. Non-perturbative masking encompasses the following techniques:
    - Sampling: this technique focuses on the release of a sample *S* of the original dataset *X*. Sampling is suitable for qualitative identifiers upon which arithmetic operations cannot be done, such as the eye colour of an individual or the months of the year. This is because sampling alone leaves a qualitative identifier unperturbed for all records in *S,* thus reducing the probability of creating unique matches. Conversely, sampling may need to be combined with other masking techniques to provide anonymous data where quantitative identifiers upon which arithmetic operations can be done are in place, such as the height or salary of an individual. This is because quantitative identifiers have more variability, therefore being more unlikely for two identifiers to take the same value.
    - Generalization: this technique focuses on the reduction of the granularity of data so that dataset *Y* is less precise than dataset *X*. This technique is more appropriate for qualitative identifiers, as it supports the disguise of records with unusual combinations. Further, this technique works better if many individuals can be subsumed within the generalized identifiers created in dataset *Y*. For instance, if a given dataset *X* holds information about

[52] According to the Irish Data Protection Commission, masking alone entails a very high risk of identification, and so will not normally be considered anonymisation in itself. See Irish Data Protection Commission, Guidance Note: Guidance on Anonymisation and Pseudonymisation, p. 12. Hence, masking may work as a supplement to other anonymization techniques. A paradigmatic example of masking is to be found in billing scenarios. The most common example includes the masking of credit card information, e.g. it is displayed in the form of XXXX XXXX XXXX 4321.

various professions, such as 'accountant', 'statesman', 'banker', 'physician', 'dentist', and 'pharmacist', a dataset *Y* may generalize said information by creating identifiers such as 'financial professionals' and 'sanitary professionals'. Generalization is used heavily by statistical offices and institutes.

- Top and bottom coding: this technique is a special case of generalization by which top-codes or bottom-codes are set from the original identifiers of dataset *X.* A top-code is the upper limit among all values of dataset *X,* and a bottom-code represents the lower limit. The idea is that top-codes which are above certain threshold are lumped together to form a new identifier. The same is done for bottom-codes which are below a certain threshold.
- Suppression: this technique focuses on the removal of the entire or certain identifiers in dataset *Y* before its release. Since the recovery of information is not possible under this technique, suppression is considered the strongest anonymization technique.[53] Different types of suppression can be differentiated. Local suppression: this technique focuses on the removal of certain individual identifiers in dataset *Y* with the aim of increasing the set of records that share a combination of key values. Local suppression is rather oriented to qualitative identifiers. Record suppression: this technique focuses on the removal of an entire record in dataset *Y*. Since record suppression affects multiple identifiers at the same time, the statistical properties of dataset *Y* can be affected with respect to dataset *X*.

- Perturbative masking: this technique focuses on the distortion or perturbation of microdata so that the statistical properties of the original dataset *X* are preserved in dataset *Y*. Perturbative masking encompasses the following techniques.
  - Noise addition: this technique focuses on the masking of identifiers in a given dataset *X* by adding random noise. Since noise addition modifies the statistical properties of dataset *Y* with respect to the original dataset *X*, the statistical properties of the noise being added determine the effect of noise addition. The amount of noise should be proportionate to the range of values of the identifiers. If the base is too small, the anonymization effect will be weaker; on the other hand, if the base is too large, the end values will be too different from the original dataset *X,* and the utility of dataset *Y* will likely be reduced. This technique is generally applied to quantitative identifiers. Since noise addition focuses on the preservation of the privacy guarantees rather than the statistical properties of the data, where accuracy is crucial, noise addition is not recommended.
  - Data swapping: this technique focuses on the transformation of a database *X* by exchanging identifiers among individual records. Ideally, the process is irreversible so that the original dataset *X* cannot be retrieved from the swapped dataset *Y*. For example, #458912 may become #298514. Data swapping is recommended where subsequent analysis only needs to look at aggregated data or where there is no need for analysis of relationships between identifiers at the record level.
  - Microaggregation: this technique focuses on the clustering of records of dataset *X* into small aggregates or groups of *k* elements, where the average of the values of the group over which the record belongs is published in dataset *Y.* The aggregates should be as homogeneous as possible to minimize information loss. To obtain microaggregates in dataset *X* with *n* records, these are combined to form groups of size at least *k*. For each identifier, the average value over each group is computed and is used to replace each of the original averaged values. For instance, given dataset *X* containing the ZIP, age, and disease of eight persons, dataset *Y* can be released by implementing microaggregation. As shown below, Dataset *Y* is an anonymized version of dataset *X,* where the privacy parameter has been set to 4. This implies that for any combination of values of indirect

[53] Personal Data Protection Commission. Singapore. Guide to basic data anonymisation techniques. 25 January 2018, p. 12.

identifiers in dataset *Y,* there are at least four records sharing that same combination of values.[54]

	ZIP	Age	Disease
x_1	35008	21	COVID -19
x_2	35007	23	COVID -19
x_3	35007	20	COVID -19
x_4	35006	18	COVID -19
x_5	30000	50	COVID -19
x_6	30000	49	COVID -19
x_7	29500	47	Pneumonia
x_8	29500	42	Pneumonia

Table 1: Dataset X

		ZIP	Age	Disease
G_1	x_1	35007	21	COVID -19
	x_2	35007	21	COVID -19
	x_3	35007	21	COVID -19
	x_4	35007	21	COVID -19
G_2	x_5	30000	47	COVID -19
	x_6	30000	47	COVID -19
	x_7	30000	47	Pneumonia
	x_8	30000	47	Pneumonia

Table 2: Dataset Y

2.3.1.1.2 Data synthesis

Data synthesis aims at creating a dataset *Y* which consists of randomly simulated records that do not directly derive from the dataset *X* while preserving the statistical properties of the original dataset *X.* As such, standard deviations, medians, linear regression, or other statistical techniques can be used to generate synthetic data. The generation of a synthetic dataset *Y* takes chiefly three steps: (*i*) the proposition of a model *y* for a given population; (*ii*) the adjustment of the proposed model *y* to the original dataset *X;* and (*iii*) the generation of a synthetic dataset *Y* drawn from the adjusted model *y*.

- Full data synthesis, where every identifier for every record has been synthesized, *i.e.* identifiers contained in dataset *Y* are a new sample from the underlying dataset *X.*
- Partial data synthesis, where only identifiers with high risk of disclosure are synthesized.
- Hybrid data synthesis, where dataset *X* is mixed with a fully synthetic dataset *Y.*

As put forward by Domingo-Ferrer et al., the utility of a synthetic dataset *Y* is highly dependent on the accuracy of the adjusted model *y*. If the adjusted model *y* fits well the population, the synthetic dataset *Y* should be as good as the original dataset *X* in terms of statistical analysis utility. As a result, synthetic data are superior in utility to other masking techniques which may be subject to a trade-off in utility. However, it should be equally noted that the proposition of a model *y* which appropriately captures all the properties of the population is, in general, a complex task.

2.3.1.2 Privacy models in microdata releases

[54] Dataset Y can however be subject to linkage attacks. For instance, if an intruder discovers via other external means some information of a given x_i, he may be able to infer some information of the data subject. Suppose that the intruder knows x_3 ,who lives in 35007 and is 20 years of age. Then, by linking this information with dataset Y, the intruder can conclude that x_3 has COVID – 19. Cf. Abidi, B., Ben Yahia, S., & Perera, C. (2020). Hybrid microaggregation for privacy preserving data mining. Journal of Ambient Intelligence and Humanized Computing, 11(1), 23-38.

A privacy model is a framework that specifies the conditions that dataset *Y* must satisfy to keep the disclosure risk of data under control.[55] Privacy models are subject to one or several parameters which determine the acceptability of risk. They can be equally nurtured by one or several anonymisation and/or pseudonymisation techniques to maintain risk under control. Due to their particular nature, privacy models tend to isolate variables which are controllable. Therefore, existing privacy models are developed under static conditions, which may not take into account big data settings. This may be a friction point in relation to the European data protection framework, by which the assessment of risk should be done from an absolute approach, *i.e.* an approach that embraces any theoretical and/or remote probability of re-identification in relation to a given data processing.

While anonymisation techniques specify the technical processing with the aim of limiting the risk of re-identification, they do not pronounce themselves about the assessment of the risk itself. For these purposes, it is the privacy model the suitable framework in charge of quantifying the risk of re-identification of the anonymized data. However, there is always a trade-off. As privacy models tend to specify the required properties that a dataset *Y* must satisfy to limit the re-identification risk, they leave it open to the controller or processor to choose which anonymization or pseudonymization technique to apply to satisfy those properties.

Different privacy models have been proposed, including k-anonymity,[56] l-diversity,[57] t-closeness,[58] and ε-differential privacy.[59] Figure 2 below is a non-exhaustive list of the most relevant privacy models as identified in the literature. The following paragraphs attempt to shed light on the main properties and implications of these privacy models in relation to the risks to re-identification.

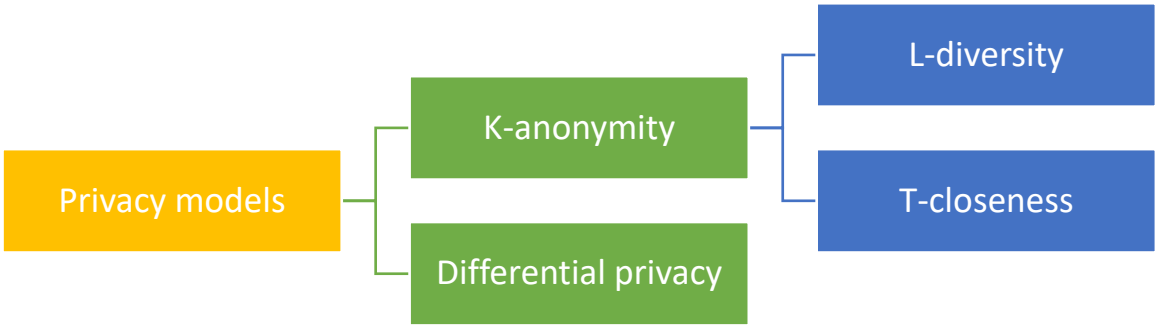

Figure 2: Privacy models

- *k*-anonymity: this model seeks to prevent the re-identification of records based on a predefined set of indirect identifiers. To this extent, *k*-anonymity departs from the assumption that a record identification is performed based on a fixed combination of identifiers, and it seeks to make said combination refer to at least to *k* individuals. For instance, given a combination of indirect identifiers corresponding to a data subject, the probability of those indirect identifiers being linked to the said data subject is 1/*k*. Since a potential attacker cannot be fully sure about the certainty of said combination, this results in the prevention of identity disclosure because the same identifiers are shared by many data subjects, *i.e.* it prevents an intruder from singling out a data subject. It should be noted, however, that while *k*-anonymity may offer appropriate solutions for re-identification, it does not prevent the disclosure of identifiers, *e.g.* where the

[55] Soria-Comas, J., & Domingo-Ferrer, J. (2016). Big data privacy: challenges to privacy principles and models. Data Science and Engineering, 1(1).
[56] Samarati P. (2001) Protecting respondents' identities in microdata release. IEEE Trans Knowl Data Eng 13(6):1010–1027.
[57] Machanavajjhala A., Kifer D., Gehrke J. & Venkitasubramaniam M. (2007). L-diversity: privacy beyond k-anonymity. ACM Trans Knowl Discov Data, 1(1):3.
[58] Li N. & Li T., Venkatasubramanian S (2007) t-Closeness: privacy beyond k-anonymity and l-diversity. In: Chirkova, R., Dogac, A., Özsu, M.T. & Sellis, TK. (eds) Proceedings of the 23rd IEEE international conference on data engineering (ICDE 2007), p 106–115.
[59] Dwork, C., McSherry, F., Nissim, K. & Smith, A. (2006) Calibrating noise to sensitivity in private data analysis. In: Halevi, S. & Rabin, T. (eds) Proceedings of the third conference on the theory of cryptography, vol 3876., lecture notes in computer science. Springer, p 265–284.

variability of identifiers in a *k*-anonymous group of records is small. Two attacks have been proposed in the literature that exploit the lack of variability of identifiers. On the one side, a background knowledge attack can be performed where the variability of identifiers in the *k*-anonymous group is small and the intruder is in possession of some background information which allows him to target the data subject. On the other side, a homogeneity attack can be performed where all the records in a *k*-anonymous group share the same identifier. In this case, an intruder can conclude that a data subject corresponds to a certain group given the indirect identifiers of the said data subject. A homogeneity attack can be better illustrated in the following tables.

	Non-sensitive identifiers			Sensitive identifiers
	Age	ZIP	Nationality	Condition
x_1	18	35008	Spanish	COVID -19
x_2	19	35006	Belgian	COVID -19
x_3	22	35008	Moroccan	Gastritis
x_4	18	35007	Spanish	COVID -19
x_5	52	37008	Spanish	Gastritis
x_6	50	37010	Spanish	Gastritis
x_7	41	37008	Finish	Cancer
x_8	57	37008	Belgian	Cancer
x_9	35	35007	Spanish	Gastritis
x_{10}	34	35001	Finish	Gastritis
x_{11}	39	35005	Belgian	Gastritis
x_{12}	37	35002	Moroccan	Gastritis

Table 3: Dataset X

	Non-sensitive identifiers			Sensitive identifiers
	Age	ZIP	Nationality	Condition
x_1	< 25	35***	***	COVID -19
x_2	< 25	35***	***	COVID -19
x_3	< 25	35***	***	Gastritis
x_4	< 25	35***	***	COVID -19
x_5	≥ 50	37***	***	Gastritis
x_6	≥ 50	37***	***	Gastritis
x_7	≥ 50	37***	***	Cancer
x_8	≥ 50	37***	***	Cancer
x_9	3*	35***	***	Gastritis
x_{10}	3*	35***	***	Gastritis
x_{11}	3*	35***	***	Gastritis
x_{12}	3*	35***	***	Gastritis

Table 4: Released dataset Y based on k-anonymity

As shown in Table 4, the released dataset *Y* satisfies the *k*-anonymity model where $k = 4$, *i.e.,* every group of x_i has a minimum size of 4. For the release of the dataset, the *4*-anonymity model has been complemented by anonymization techniques such as the generalization of the identifier age, the masking of the ZIP codes, and the suppression of the nationality. However, the third group encompassing $x_9 - x_{12}$ contains the same condition for each x_i. Therefore, a potential attacker knowing an indirect identifier of a data subject, *e.g.,* the age of a data subject, may perform a homogeneity attack and conclude that the data subject belongs to that group and hence suffers from gastritis. In the background knowledge attack scenario, the intruder has some additional knowledge about the data subject which he has acquired from an external source of information. According to this setting, if an attacker knows, for instance, that a data subject

belongs to the first group $x_1 - x_4$ and does not have gastritis, he may conclude that the data subject suffers from COVID-19.[60] In order to avoid the previous pitfalls, *k*-anonymity is usually complemented with further models and techniques aiming at protecting the disclosure of identifiers to prevent the gain of knowledge about the data subject. These include, inter alia, the privacy models of *l*-diversity and *t*-closeness.

- *l*-diversity: the goal of this model is to require a minimum level of diversity for the identifiers that may be sensitive or compromising to a data subject in each of the *k*-anonymous groups of records. In other words, *l*-diversity aims at ensuring that each group of sensitive identifiers contains different values and that none of these values dominates in terms of frequency of appearance. Although several diversity parameters have been proposed, the simplest notion of *l*-diversity considers that there should be at least *l* different values for each sensitive identifier in order to have robust prevention against the gaining of knowledge by an intruder about a given data subject. However, *l*-diversity has several drawbacks, as it is not a resilient privacy model against skewness and similarity attacks.[61] Skewness attacks occur when the distribution of values of the sensitive identifiers within a given group is different from the distribution of values for the same identifier over the total population, *e.g.,* if a medical dataset *X* contains a rare disease for which only 1% of the population is positive is released under the *l*-diversity model, and the anonymized dataset *Y* configures the population as having 50% probabilities of being positive, the released dataset *Y* will be skewed or distorted in comparison with the original dataset *X,* for which 99% of the data subject are negative. Therefore, a potential intruder knowing that a certain data subject is contained in the dataset *Y* may erroneously gain sensitive information about the said data subject. Similarity attacks occur when the values of sensitive identifiers are different but semantically similar. In this case, an attacker can increase his or her knowledge about the value of the sensitive identifier associated with the data subject within the group in which the data subject is included. For instance, if the data subjects have been classified in a *l*-diverse group in dataset *Y* according to three different types of influenza or flu, *e.g.* influenza A, B, and C, an attacker can conclude that a data subject meeting the identifiers of that group has influenza.
- *t*-closeness: the goal of this model is to overcome the disadvantages of the *l*-diversity model. To this extent, *t*-closeness proposes the use of a relative tool to measure the variability of the values of the sensitive identifiers, thus limiting the information gain about the data subjects. In the *t*-closeness model, all values assumed by the sensitive attribute are considered equally sensitive. The model requires the value distribution of the sensitive identifiers in each group to be close to the value distribution in the released dataset *Y*, where the difference among them is lower than the threshold *t.* As a result, skewness attacks are not possible since the *t*-closeness model reduces the difference between the value distribution of the sensitive identifiers in each group and population. Further, similarity attacks become more challenging due to the fact that the semantical similarities among identifiers in each group do not provide additional information with respect to the whole dataset *Y.*[62] Consider the following table, where the *t*-closeness model has been applied.

Birth date	Sex	ZIP	Disease
1993	F	35***	COVID-19
1993	F	35***	Gastritis

[60] Bijl, A. F. (2017). Data Anonymisation in the light of the General Data Protection Regulation (Doctoral dissertation, Faculty of Science and Engineering), p. 11.
[61] Li, N., Li, T., & Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k-anonymity and l-diversity. In 2007 IEEE 23rd international conference on data engineering (pp. 106-115). IEEE.
[62] Garcia-Alfaro, J., Navarro-Arribas, G., Cuppens-Boulahia, N., & Roudier, Y. (2011). Data privacy management and autonomous spontaneous security. In Proceedings of 5th International Workshop, Dpm 2010 and 3rd International Workshop, SETOP (Vol. 5).

1993	F	35***	Leukemia
1994	M	35***	Leukemia
1994	M	35***	Cancer
1994	M	35***	COVID-19
1993	F	35***	Hypertension
1993	F	35***	COVID-19
1993	F	35***	Gastritis

Table 5: An example of a 3-diverse table based on the t-closeness model

- As shown above, in each one of the 3-anonymous groups, the sensitive identifiers, i.e., the ones displaying the disease, preserve distribution equality. They are different and not systematically similar, therefore rendering void the application of any skewness or similarity attack. However, one of the undesired effects of this model is the decrease in the relationship between indirect identifiers and their sensitive values. For instance, if a certain group contains a sensitive value which does not appear in the overall distribution, t-closeness model would mandate to generalize or suppress the value, thus decreasing the utility of the dataset Y. Although an increase of the parameter t may solve this issue, thus will likely result in a higher vulnerability to similarity attacks.

- ε-differential privacy: this model aims to protect privacy in interactive settings by controlling the release of information of queries to a database *Y*. Differential privacy (DP) has emerged as an anonymization technique in computer science that allows accurate data mining and sharing while preserving formal privacy guarantees.[63] DP is defined as a mathematical definition of privacy in the context of statistical and machine learning analysis.[64] DP guarantees, in a mathematical sense, that the pair of outputs produced by two neighbouring databases (which are the same except for one user's data) are nearly indistinguishable. This means that the inferences that can be made from a differentially private analysis are essentially equal, whether or not said individual's private information is included in the input to the analysis. DP typically works by adding some noise to the data. The amount of noise added is determined by a privacy loss parameter, which is usually denoted by the Greek letter ε (epsilon), as illustrated in the figure below. Differential privacy has gained a lot of attention due to its robust privacy guarantees as well as its advantages to data utility preservation. The model was first proposed by Dwork and assumes anonymization as a mechanism that preserves the knowledge gain derived from the presence of an individual in a dataset. In this way, the presence or absence of any single individual record in the database or dataset should be unnoticeable when looking at the responses returned for the queries. An algorithm that satisfies differential privacy has its input in the original dataset *X* of the data controller and produces different datasets *Y* that differ in one single record while preserving the statistical properties among the released datasets.
  As its output, it produces tables that differ in one single record. The property of differential privacy is that the probabilities of these different tables will be almost similar. The difference between these tables is notated as ε, which is a parameter that controls the amount of random noise added to the response to each query. This noise addition has the property that an attacker cannot find the differences of the datasets as a consequence of the increased fluctuation in relation to the difference of one record. As a result, an attacker cannot learn the data of a data subject since he or she cannot differentiate between two data sets that differ in one person. One of the drawbacks of this model is its vulnerability to counting attacks, where an attacker queries the same dataset *Y* multiple times in order to draw sound conclusions about the initial dataset *X*. Consequently, a robust differential privacy model must keep track and set a limit on the number of particular queries. Further, the setting of the value ε is a problematic issue, as it leads to a trade-off between privacy and utility.

[63] Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In Theory of cryptography conference (pp. 265-284). Springer, Berlin, Heidelberg.
[64] Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., Honaker, J., ... & Vadhan, S. (2018). Differential privacy: A primer for a non-technical audience. Vand. J. Ent. & Tech. L., 21, 209.

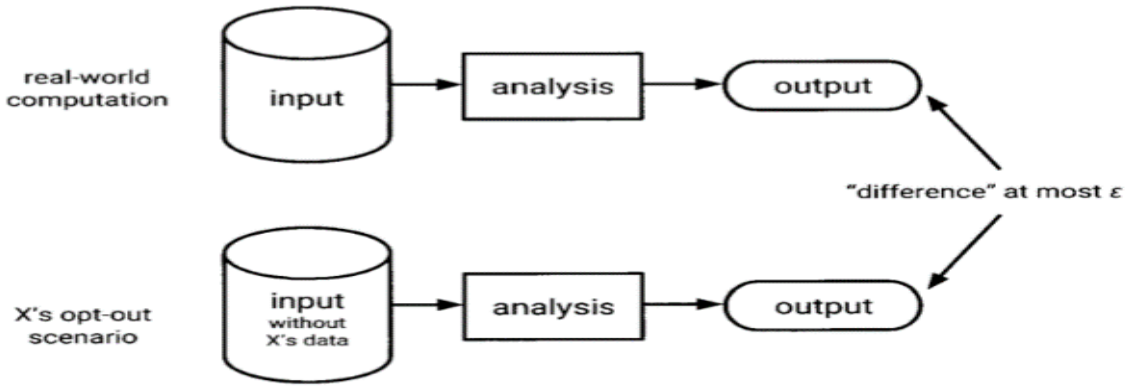

Figure 3: Differential privacy[65]

Noise introduced masks the differences between the real-world computation and the opt-out scenario of each individual in the dataset, turning the outcome of a differentially private analysis into an approximation. This implies that, if a differentially private analysis is performed twice on the same dataset, the result will intentionally differ due to the addition of said random noise. However, for large enough datasets and with the information on the noise-adding mechanism, the dataset is still accurate in terms of aggregate measurements. The amount of noise added is hence a trade-off, as adding more noise makes the data more anonymous, but it also makes the data less accurate. Therefore, small ε is associated with stronger privacy guarantees but weaker accuracy. Contrarily, big ε is associated with weaker privacy guarantees but stronger accuracy. Much of the debate in literature advocates setting ε to be a small constant $\varepsilon < 1$, or to be diminishing in the size of the database for a database of size n.[66]DP can be implemented locally or globally. In local settings, noise is added to individual data before its centralization in a database. In global settings, noise is added to raw data after its collection by a trusted third party or curator. Moreover, different DP mechanisms can be used for different analytical tasks, for instance, for generating a machine learning model,[67] releasing micro-data,[68] or building a histogram.[69] DP has several advantages:

- One of the advantages of DP is its deniability aspect. As previously introduced, each query to a database where DP has been applied would lead to a different answer. These approximately similar answers are still meaningful for aggregate statistics. However, the querier cannot reveal the specific information about the individuals contained in the database. This deniability is an important feature against, inter alia, linkage attacks where attackers leverage multiple sources to identify the personal information of a target. One paramount example of linkage attacks is the one performed by Latanya Sweeney, who was able to reveal Massachusetts Governor William Weld's medical records by combining anonymized public data from an insurance agency for state employees and voter registration records that were publicly available for a small fee.[70] In this way, while acknowledging its intrinsic limitations,[71] DP offers better chances to prevent re-identification as opposed to traditional anonymization techniques such as removing columns containing personally identifiable information or data masking.

[65] ibid., 235.
[66] Hsu, J., Gaboardi, M., Haeberlen, A., Khanna, S., Narayan, A., Pierce, B. C., & Roth, A. (2014). Differential privacy: An economic method for choosing epsilon. In 2014 IEEE 27th Computer Security Foundations Symposium (pp. 398-410). IEEE.
[67] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security (pp. 308-318).
[68] Bild, R., Kuhn, K.A., & Prasser, F. (2018). SafePub: A Truthful Data Anonymization Algorithm With Strong Privacy Guarantees. Proceedings on Privacy Enhancing Technologies, 2018, 67 - 87.
[69] Dwork C. (2008) Differential Privacy: A Survey of Results. In: Agrawal M., Du D., Duan Z., Li A. (eds) Theory and Applications of Models of Computation. TAMC 2008. Lecture Notes in Computer Science, vol 4978. Springer, Berlin, Heidelberg.
[70] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.
[71] Holzel, J. (2019). Differential Privacy and the GDPR. Eur. Data Prot. L. Rev., 5, 184.

- Furthermore, DP enables the customization of privacy levels by the adjustment of the parameter ε, thus controlling the level of privacy according to, inter alia, the sensitivity of the dataset. This is not the case for other privacy-enhancing technologies which are not process-based, for which the calibration of noise to the query is not possible, such as masking or generalization. DP helps hence to navigate the trade-off between privacy and utility.

Contrarily to the afore-mentioned benefits of DP, several drawbacks must also be highlighted.

- First, DP requires restricting the questions to aggregates. This is a requirement stemming from the necessary addition of noise, which can only be done on numerical values. Hence, literal data cannot be subsumed under DP models.
- Second, the number of queries in the dataset is a point of concern, as they are intimately related to the loss of privacy. Hypothetically, with each query, a third party could use their aggregate results to reconstruct the original data by filtering out the noise through averaging. This could put at risk the identifiability of the data subjects.
- Third, the trusted model followed can equally determine the effectiveness of DP. For instance, global DP models assume a curator who has full access to the unprotected database as well as calculates the necessary perturbations to be added to the query results. This is not an optimal approach to data protection, as the general data protection framework does not recognize the trustworthiness of any party holding personal data. In order to remedy the restrictions of centralized models, local models of DP have been proposed, where data subjects themselves generate differentially private query results.
- Other drawbacks of DP are related to the size of the dataset upon which differential computation is applied. Whereas DP techniques can be ostensibly introduced in large datasets without compromising their accuracy, the noise added in small datasets can seriously impact their analysis. Lastly, the lack of consensus surrounding the optimal value of ε is also a concern widely addressed in the literature.[72]
- The most disputed problem in differential privacy may be the suitability of this technique as an anonymization technique in relation to the legal definition of anonymization. It has been proposed that, contrarily to protecting data subjects fundamental rights and freedoms in a general way, DP aims at reducing the chance that a data subject faces any harm which is specific to their participation in a statistical database, but not as to their existence. This results in DP enabling the limitation of the knowledge gain of an attacker on individuals in cases where their participation would not make a significant statistical difference, but it does not prevent record linkage by an attacker. For example, Hölzel proposes that DP only assures that any distinct input to a query does not have a meaningful influence on the query result. On the other hand, for example, Cohen and Nissim have a more narrow and technical approach towards DP.[73] According to Cohen and Nissim, the 'singling out' threshold contemplated by the GDPR should be understood as a type of privacy attack intended to capture that concept, *i.e.,* as an attack where an adversary predicate singles out a dataset *x* using the output of a data-release mechanism *M(x)* leading to the finding of a predicate *p* matching exactly one row in *x* with probability much better than a statistical baseline. DP can preclude this type of attack and would categorize as an effective anonymization technique. This, in turn, should be evaluated as a mathematical concept with the legal consequence of rendering personal data non-identifiable in line with the definition of personal data in the GDPR. By leveraging a connection to statistical generalization, the authors show that DP can prevent such attacks, therefore categorizing it as an anonymization technique.[74] More

[72] Nozari, E., Tallapragada, P., & Cortés, J. (2015). Differentially Private Average Consensus with Optimal Noise Selection. IFAC-PapersOnLine, 48, 203-208.
[73] Cohen, A., & Nissim, K. (2020). Towards formalizing the GDPR's notion of singling out. Proceedings of the National Academy of Sciences of the United States of America, 117(15), 8344–8352.
[74] It should be noted that the authors acknowledge that the prevention of singling out attacks in a dataset are a necessary (but maybe not sufficient) precondition for a dataset to be considered effectively anonymized.

concretely, the authors postulate that, since a predicate singling out attack implies a form of overfitting to the underlying dataset, DP mechanisms can reasonably prevent this form of overfitting and, hence, protects against predicate singling out, as opposed to, for instance, *k*-anonymity.[75] Such an assumption should be contrasted with the approach of the Article 29 Working Party, for which *k*-anonymity, singling out is no longer a risk, whereas with differential privacy, it 'may not' be a risk.

2.3.2 Results from interviews

This section will discuss some of the main findings gained through the interviews conducted for this study. The full interview reports may be found in the annex to this report.

In terms of distinguishing between anonymization techniques and pseudonymization techniques, interviewees indicate that this is not a distinction that is easily made, if at all. While the legal framework distinguishes between the two modalities, on a technical level, it is not so much an either/or distinction. Thus, it is also not that easy to say if a technique is an anonymization technique or just a pseudonymization technique. Technical experts are reluctant to use the term anonymous/anonymity in the way it is understood in the GDPR.

From a technical perspective, data could be called anonymous data when a number of relevant variables are removed. In addition, it should be emphasized that it is not so much the technique or technology that determines whether a process is anonymisation or pseudonymisation. Techniques can be used in different ways. Rather, one should look at the purpose of the processing. Many technical experts assume levels of anonymity, for example, partial or full anonymity. There is a scale from full anonymity to direct identifiability rather than a binary distinction, as is prevalent in the GDPR. In addition, the fact that the GDPR sets no time limit on when data can be re-identified or de-anonymised means that, keeping an eye on the technological developments, it is highly likely that at some point in time, the data will be considered personal data again.

There are misconceptions within the technical community with respect to the legal definition of anonymity, e.g., some actors claim to anonymize data but, in reality, are merely removing some identifiers or are pseudonymising the data from a legal perspective. A number of technical experts question whetherthe legal definition of anonymous data can be upheld, as it will be increasingly difficult to meet the legal threshold. From their perspective, it is almost impossible to truly have anonymous data. In particular, when anonymised datasets are shared or made available online, it is likely that there will be a party that re-identifies the data or merges it with other datasets to arrive at personal data. That is why some experts speak of presumed anonymous data. With synthetic data, 'real data' is mixed with 'fake data'. This could be a way to arrive at anonymous datasets, but it also entails the risk that fake data are attributed to real people.

2.3.3 Results from workshop

The workshop held for this study yielded the following results:

- Complexity of the terminology: a general sentiment that was shared is that the term anonymisation was unclear and vague due to its many open-ended factors not only in the legal text but also from the different technical concepts. A special point of reference was the term 'reasonably likely'.
- Black or white approach: the black or white approach of anonymity under the GDPR can be a demotivating factor for data controllers/processors, as it can be difficult to achieve true

[75] K-anonymity can enable an adversary to predicate single out with probability approximately 37%, even using extremely low-weight predicates for which the baseline risk is negligible.

anonymity. There is a misalliance between the technical/mathematical perspective, in which full anonymization is almost never possible, and the legal regime.

- Contextual norms: it could be considered to take account of the differences between various actors. Different actors have different resources in terms of anonymizing and de-anonymizing data. These differences could be taken into account by the addressees of norms.
- Differentiation: the legal regime makes no difference between anonymous data and aggregated data, while from a technical perspective, the difference is significant. On a record level, it is almost impossible to speak of anonymous data, whereas on aggregate level, there are many more opportunities to protect individuals from identification.
- Availability of data: differential privacy and k-anonymity have failed in providing absolute protection so far. The main reason is that other types of information and datasets are available, often online. The effects thereof are difficult to estimate when realising or sharing data.
- Location data: location data is particularly challenging to anonymize.
- A granular approach: an option could be to apply the GDPR in a contextual way; that is: the more data can be considered anonymous, the less stringent the GDPR obligations will be applied.

2.4 Analysis

Four main tensions between the legal and the technical realm have emerged from this chapter:

1. A major challenge for data protection is the anonymization of personal data, more specifically, the problem of achieving true anonymity.[76] Out of all the legal concepts discussed in this report, the concept of anonymous data is perhaps the most contested. The main criticism is the disconnect between the legal terminology or scope and the technical reality. On the one hand, from a technical point of view, the legal concept is conceived as lacking rigor and clear definition.[77] This tension raises the question of whether anonymization is the correct term or the threshold that regulation should strive for. On the other hand, in information theory, anonymity is much more strictly defined, meaning that anonymous data will in principle, be valueless. From a technical perspective, it is clear that complete anonymization is impossible to achieve given the legal standard set out by data protection law and the contextual nature of information.[78] That is why experts generally propose to develop a framework that minimizes the risk of re-identification.[79] In the same way that locking the doors and windows to one's home reduces the risk of unwanted entry but is not 100% safe, so too anonymization should be understood.[80] It is also clear that the general availability of data makes it increasingly unlikely that anonymous datasets that are shared or made available stay anonymous.
2. Then there are different perspectives on the value and protection of data. The legal framework is based on the assumption that personal data have a higher value to the data subject than anonymised data, in the sense that personal data would relate more to the data subject, and therefore higher protection should be granted to safeguard its fundamental rights, such as privacy. However, from a technical point of view, aggregate or anonymized individual data can be valuable as well, for example, for predictive analytics or for constructing group profiles. Interviews with technical experts demonstrated that actors could derive attributes or information from anonymous datasets and use those without knowing the identity of the person. Thus, if the use of anonymous data is not without potential harm, this challenges the presumption that

[76] Ohm, P. (2010). Broken Promises of Privacy, 57 UCLA L. Rev, 1701, 1719. Schwartz, P. M., & Solove, D. J. (2011). The PII problem: Privacy and a new concept of personally identifiable information. NYUL rev., 86, 1814.
[77] Nissim, K. (2021). Privacy: From database reconstruction to legal theorems. In Proceedings of the 40th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (pp. 33-41).
[78] IAPP, Looking to Comply with GDPR? Here's a Primer on Anonymization and Pseudonymization, Apr. 27, 2019.
[79] El Emam, K. (2013). Guide to the de-identification of personal health information. CRC Press.
[80] Ontario. Office of the Information and Privacy Commissioner, Cavoukian, A., & Castro, D. (2014). Big data and innovation, setting the record straight: de-identification does work. Information and Privacy Commissioner, Ontario.

underlies the choice not to regulate anonymous data and can be a factor to consider in also subjecting anonymous data to regulation.

3. Some authors conclude that the state of the art linked to the techniques listed by Article 29 Working Party confirms that anonymization methods face big challenges with real data and that it can no longer be considered from a static perspective, but only from the dynamic one, being a dynamic checking process. Podda and Palmirani researched the various anonymization techniques and concluded that the techniques proposed by Article 29Working Party are outdated, and due to the available technology and continuous development, it is recognized that the simple model of anonymization is unrealistic.[81] According to them, research should focus on exploring new models of anonymization, such as combining techniques. An example of this is combining many techniques in a pipeline while at the same time keeping them monitored over time in a process capable of also providing a dashboard where the human expert remains in the loop.
4. What stands out from a legal perspective is the ambiguous choices made by the EU regulator. On the one hand, it keeps a strict and binary distinction between personal data and anonymous data and has adopted a regulation with respect to non-personal data, which is the mirror image of the GDPR. On the other hand, it has introduced numerous contextual elements both in the definition of personal data and in the description of anonymisation, making the assessment of whether data is personal or not increasingly fluid and contextual. Both approaches, the binary and the contextual, have raised criticism from technical experts, the first for being unrealistic and out of touch with the more fluid technological reality, the second for being vague and difficult to grasp.

[81] Podda, E., & Palmirani, M. (2020). Inferring the Meaning of Non-personal, Anonymized, and Anonymous Data. In AI Approaches to the Complexity of Legal Systems XI-XII (pp. 269-282). Springer, Cham; see also Jakob, C.E.M., Kohlmayer, F., Meurers, T., Vehreschild, J.J., Prasser, F.: Design and evaluation of a data anonymization pipeline to promote Open Science on COVID-19. Sci. Data 7, Article no. 435 (2020).

Chapter 3: Aggregation and composition

3.1 Introduction

This chapter describes the boundary between personal data and aggregated data. While the previous chapter concerned the possibility of stripping the data, still concerning one individual or a small group, of identifiable data, this chapter will discuss the process of aggregating data. Through aggregation, data can be rendered anonymous by subsuming it into higher order sets. In this way, the data is no longer treated at individual level but at group level, therefore preventing re-identification. (e.g., n > 100). This chapter will discuss different techniques for aggregation and de-aggregation, the special position of statistical data in the data protection regime, and the rules that apply to the processing of statistical data. Section 3.2 discusses the legal distinction between aggregated and non-aggregated data, section 3.3 describes the main techniques available for aggregating and re-identifying data and section 3.4 analyses the gap between the legal regulation and technical reality.

3.2 Legal regulation

This section will give an overview of the legislative history of European data protection law as far as relevant for aggregate data (section 3.2.1) and the main outlines of the rules on open data and re-use of public sector information as proposed by the EU, and its conflicts with data protection law (section 3.2.2).

3.2.1 Data Protection law

3.2.1.1 Resolution 1973

The special status of statistical and aggregate information has been part of data protection law since its beginning. In the Resolution on the protection of the privacy of individuals vis-à-vis electronic data banks in the private sector from 1973, by the Council of Europe committee of ministers, article 10 stressed: 'Statistical data should be released only in aggregate form and in such a way that it is impossible to link the information to a particular person.' Though, in essence, it may almost be regarded a tautology, because statistical data are almost always aggregate data, this provision makes clear that if statistical information can be used to identify persons, the data protection regime applies. The explanatory report makes clear that one 'of the main purposes of the data bank is to provide managers with statistical information, which will enable them to make executive decisions. Thus, the production of statistical information from data banks is a common practice. Normally, statistical data are diffused in published form. However, computerised statistics may also be made available unpublished, for example, by transfer of tapes. Owing to the special facility of computers to trace correlations, the latter form of diffusion of statistical data may also create certain dangers to privacy. The word "released" covers all forms of diffusion.'[82]

3.2.1.2 Resolution 1974

The subsequent Resolution on the public sector from 1974 included a reformulation of the principle. Interestingly, the explanatory report not only referred to the dangers of privacy, but also of discrimination with regard to processing, publishing, and using statistical information. Again, like the private sector, the report made clear that perhaps the primary reason for having databanks in the public sector was to provide governmental organisations with statistical information on which they may base

[82] <https://www.legislationline.org/documents/id/6498>.

their decisions. Interestingly, the notion of open access was also highlighted, as well as a distinction between macro- and microdata. 'Statistical information should normally be released only in aggregate form. If person by person information is released, for example for scientific or research purposes, it should be reduced to a level where it is impossible to identify the individuals.'[83]

Resolution 1974 also provided special status for statistical and aggregate data on two additional points. It made clear that it was possible to adopt special rules on the storage limitation principle if the use of the information for statistical, scientific, or historical purposes requires its conservation for an indefinite duration. 'In that case, precautions should be taken to ensure that the privacy of the individuals concerned will not be prejudiced.' The explanatory memorandum makes clear that in that case, 'data should be preserved in such a way that the identities of the people on whom information is stored can only be ascertained by the specialists carrying out the research envisaged or, in the case of other people, after an adequate period of time has elapsed.'[84] This explanation seems to foreshadow concepts such as encryption and pseudonymization. The Resolution, in the explanatory memorandum, also laid down a special status for statistical data in relation to the data quality principle. 'It was recognised that it may be impracticable or uneconomic to maintain statistical information to near perfect accuracy and to keep it absolutely up-to-date. In so far as information is provided by the individuals who are the subject of the information the accuracy of such information depends on the individuals themselves and it generally makes little difference to an individual if statistical records relating to him are not entirely accurate or up-to-date. It should also be borne in mind that when the purpose of the system is to analyse a certain set of facts, there will be no question of updating.'[85] For statistical information, it is not always important that specific information is correct, as long as the bigger picture that emerges is. Statistical information is not used to say something particular about specific individuals but to derive general and probabilistic information about groups or categories. In addition, statistics are often used for making comparative longitudinal profiles over time. Outdated information is necessary for such profiles per sé.

3.2.1.3 Convention 108

Remarkably, in Convention 108, these exceptions for processing statistical data were removed. There is no exemption for the data storage or the data quality principle, nor did the Convention contain a general rule on publishing or making available aggregated data. The Convention did contain a rule on data subjects' rights to information and rectification, for which states could adopt an exemption in relation to data processing for statistical and scientific purposes.[86] The reason for not allowing data subjects' rights to be curtailed in this context is, inter alia, that for both scientific and statistical research, it is important to keep a record of the data (correct or incorrect) which were used for research. Interestingly though, both the Convention itself and the explanatory report provide an important caveat. The restrictions to the rights may only be adopted 'when there is obviously no risk of an infringement of the privacy of the data subjects. 'Examples are the use of data for statistical work, in so far as these data are presented in aggregate form and stripped of their identifiers. Similarly, and in conformity with a recommendation of the European Science Foundation, scientific research is included in this category.'[87] Like with the Resolutions, when organisations process aggregate data or data that are stripped of identifiers, the Convention would no longer apply. This would make the exception to the data subjects' rights null and void. Convention 108+, in the wake of the EU's DPD and GDPR, also contains a special rule with respect to the purpose limitation principle for further processing for archiving purposes in the public interest, scientific or historical research purposes, or statistical purposes.

[83] <https://www.legislationline.org/documents/id/6499>.
[84] <https://www.legislationline.org/documents/id/6499>.
[85] <https://www.legislationline.org/documents/id/6499>.
[86] Convention 108, Article 9 paragraph 3.
[87] Convention 108, Explanatory Memorandum.

3.2.1.4 Data Protection Directive

The initial proposal for the Data Protection Directive of the EU only contained a limited reference to the special status of statistical processing. It stressed that countries could adopt limitations to the right to access by data subjects when their data were ´compiled temporarily for the purpose of extracting statistical information therefrom´.[88] It thus suggested dealing with the problem of many statistical agencies, namely that if they gather personal data, with the only purpose of immediately stripping the data from identities and aggregating them, the data protection framework would still apply. In addition, the explanatory memorandum specified a special status with respect to the storage limitation principle, but both this memorandum and the article in the proposed text made explicit that such should concern anonymised data only. The EU parliament suggested including a reference to the processing of statistical data on a number of points. Inter alia, it suggested an article in which public sector organisations were only allowed to communicate personal data on a limited number of grounds, one of which was when this was deemed necessary in light of statistical and scientific research. Interestingly, in its subsequent amended proposal, the Commission included a second indent in the definition of personal data contained in Article 2 a holding: ´Data presented in statistical form, which is of such a type that the persons concerned can no longer be reasonably identified, are not considered as personal data´.[89] This addition was, however, deleted from the text, as was the principle of temporarily storing data for statistical purposes and the rule on the legitimacy of communicating personal data. On the suggestion of Parliament, the final version of the Directive also included an explicit provision on the data storage principle; the Member States could adopt rules on this principle in relation to data processing for historical, statistical, and research purposes. Later on, a special status with respect to the purpose limitation principle was also included, and it was made clear that sensitive data could also be processed for statistical and research purposes.[90] Thus, the Directive restored the special status for statistical data that was contained in the CoE's Resolution from 1974.[91]

3.2.1.5 Working Party 29

The Article 29 Working Party found it important to stress that aggregate data can only be considered anonymous if the raw (underlying) data is deleted. 'For example: if an organisation collects data on individual travel movements, the individual travel patterns at event level would still qualify as personal data for any party, as long as the data controller (or any other party) still has access to the original raw data, even if direct identifiers have been removed from the set provided to third parties. But if the data controller would delete the raw data, and only provide aggregate statistics to third parties on a high level, such as 'on Mondays on trajectory X there are 160% more passengers than on Tuesdays', that would qualify as anonymous data.'[92] It also addressed specifically the promises and potential pitfalls of k-anonymity. Obviously, it warned for too low thresholds (e.g., k=2) and for inference attacks. For example, when a dataset of 100 people only specifies the year of birth and whether a person has a heart attack or not and a hacker knows a person born in 1964 to be included in that dataset, she may learn that all people born in 1964 included in that dataset had a heart attack. In a later opinion, the Working Party

[88] COM(90) 314 final ~.sYN 287 and 288 Brussels, 13 September 1990.
[89] Amended proposal for a Council Directive on the protection of individuals with regard to the processing of personal data and on the free movement of such data (*) (92 / C 311 / 04 COM (92) 422 final — SYN 287 (submitted by the Commission on 16 October 1992, pursuant to Article 149 (3) of the EEC Treaty) (*) OJ No L 111 , 5.11.1990, p. 3.
[90] See Data Protection Directive 1995, Recitals 23, 29, 34 and 40; Articles 6,11 and 13.
[91] See for interesting discussions with respect to the data protection framework for the law enforcement directive, inter alia: P6_TA(2007)0230 Protection of personal data European Parliament legislative resolution of 7 June 2007 on the proposal for a Council Framework Decision on the protection of personal data processed in the framework of police and judicial cooperation in criminal matters (renewed consultation) (7315/2007 — C6 0115/2007 — 2005/0202(CNS)). See also: P7_TA(2014)0121 European Union Agency for Law Enforcement Cooperation and Training (Europol) ***I European Parliament legislative resolution of 25 February 2014 on the proposal for a regulation of the European Parliament and of the Council on the European Union Agency for Law Enforcement Cooperation and Training (Europol) and repealing Decisions 2009/371/JHA and 2005/681/JHA (COM(2013)0173 — C7-0094/2013 — 2013/0091(COD)) P7_TC1-COD(2013)0091 Position of the European Parliament adopted at first reading on 25 February 2014 with a view to the adoption of Regulation (EU) No …/2014 of the European Parliament and of the Council on the establishment of the European Union Agency for Law Enforcement Cooperation and Training (Europol) and repealing Decisions Council Decision 2009/371/JHA and 2005/681/JHA [Am. 1] *OJ C 285, 29.8.2017, p. 288–347.*
[92] <https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf>.

warned about the possibility of combining various datasets that by themselves could not be used for identification purposes, but combined could.[93]

3.2.1.6 Legislative process of the GDPR

In the GDPR, finally, much has stayed the same compared to the Directive, but in addition to the principles contained in the Directive, it also specifies, in a number of provisions, special regimes. The DPD had a special regime for the freedom of speech, stressing that the Member States could provide for exemptions or derogations from the provisions for the processing of personal data carried out solely for journalistic purposes or the purpose of artistic or literary expression (freedom of expression).[94] This included making available governmental data, which was regarded essential for governmental transparency. Moreover, the ECtHR does not only guarantee the right to impart but also to receive relevant information.[95] Interestingly, the GDPR mentions several different regimes that the Member States may adopt. Article 85 lays down a rule similar to that contained in the DPD. In addition, Article 86 holds that personal data in public sector documents may be disclosed when laid down in law in order to ensure public access to government documents. Finally, Article 89 provides for derogations for the processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes.

3.2.2 Open Data and Re-use of data

3.2.2.1 Evolution from open government to re-use of public sector information

Western society has been based on the principle of open and transparent government for centuries. The idea is that critical citizens and journalists should be able to check decision-making processes in order to expose potential problems and abuse of power. This principle also allows historians and scientists to examine archives in order to describe and verify how governments operated in certain periods. An open government is considered quintessential for a vital democracy. In this way, four important developments have taken place in recent years:

- The first is digitisation. Government documents used to be available in archives, libraries, or specially designated information centres. Nowadays, more and more documents are made available online. This has an important effect on what is called 'practical obscurity'. The fact that in the past one had to make the effort to go to the place where the documents were stored, request them, and view them meant that, in practice, only a limited number of people were able to access the information. Broadly speaking, these were journalists, historians, critical citizens closely following the government, and lay historians researching their family trees. By making the documents public on the Internet and not setting any access barriers, now anyone can view these documents with ease.
- Second, in the pre-digital age, most documents were 'passively disclosed'; citizens, journalists, and others were given access to specific documents upon request. They already had to have a rough idea of what they were looking for, the disclosure of documents required their initiative, and the documents were usually made available for a certain period of time only. Currently, documents are increasingly disclosed actively; the government publishes documents not upon request, but on its own initiative. This means that there is no longer a specific reason for which a document is made available. Anyone may access them, and they are made available permanently.
- Third, the technical possibilities of searching through such documents have increased considerably. These include algorithms and AI/tools that can analyse texts for words,

[93] <https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf>.
[94] Article 9 Data Protection Directive.
[95] <https://echr.coe.int/Documents/Guide_Art_10_ENG.pdf>.

correlations, and topics. Whereas previously, it was primarily individuals that sought access to government documents, currently, it is tech companies that are best placed to scan and analyse the millions of governmental documents that appear on the internet every year.

- Fourth, and perhaps most importantly, the European Union has encouraged the Member States to not only make data available to further open, transparent, and accountable government but also to facilitate the reuse of government data. The idea is that the government is sitting on 'a mountain of data', while its economic potential is not being exploited. Already in the year 2000, the total value of the European public sector information (PSI) was estimated to be around 68 billion euro annually.[96] The data are 'only' used for furthering public interests, while if the data were released for the commercial re-use, it is estimated that tens of billions in economic potential would be released. The EU, therefore, adopted a Directive on the re-use of public sector information in 2003, which,[97] following amendments in 2013[98] and 2019,[99] has become even more adamant that governments actively release public sector information to enable re-use by commercial parties.

3.2.2.2 Rules that apply to statistical agencies

For statistical agencies, making public general/aggregate information and allowing researchers access to specific microdata, there are a number of European-wide principles to take into account. Statistical agencies should have a legal mandate to collect and access information from multiple data sources for the development, production and dissemination of European Statistics. They have to ensure the quality, objectivity, and neutrality of the statistics and, at the same time, ensure the confidentiality and privacy of citizens. Employees have to be under strict confidentiality obligations, and when third parties want to have access to microdata for research purposes, they should follow strict protocols.[100]

3.2.2.3 Tensions between open data/re-use PSI and data protection

The European Data Protection Supervisor, when advising on a proposal for a regulation of the European Parliament and the Council on Community statistics on public health and health and safety at work, pointed out that, although 'confidentiality' and 'privacy' or 'personal data' use the same vocabulary, there are important differences. 'For instance, the definition of confidentiality also deals with non-natural/physical persons, while the notion of personal data relates exclusively to natural persons. Moreover, the definition of confidentiality, unlike the notion of personal data, excludes data taken from sources which are available to the public and remain available to the public. Therefore, some data which may not be considered as confidential anymore from a statistical point of view could still be considered personal data from a data protection point of view. The same analysis occurs with the notion of anonymity. Although from a data protection view, the notion of anonymity would cover data that are no longer identifiable (see recital 26 of the Directive), from a statistical point of view, anonymous data are data for which no direct identification is possible. This definition implies that indirect identification of data would still qualify these data as anonymous, from a statistical point of view.'[101]

In relation to the re-use of government information, there are many dilemmas concerning its legitimacy and desirability. The question is whether the reuse of governmental information for commercial ends

[96] European Commission, Commercial Exploitation of Europe's Public Sector Information, 20 September 2000, p. 6.
[97] Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information. Official Journal L 345 , 31/12/2003.
[98] Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the Re-Use of Public Sector Information.
[99] Directive 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information.
[100] <https://ec.europa.eu/eurostat/documents/4031688/8971242/KS-02-18-142-EN-N.pdf/e7f85f07-91db-4312-8118-f729c75878c7?t=1528447068000>.
[101] Opinion of the European Data Protection Supervisor on the proposal for a Regulation of the European Parliament and of the Council on Community statistics on public health and health and safety at work (COM(2007) 46 final), *OJ C 295, 7.12.2007, p. 1–6. See also:* Opinion of the European Data Protection Supervisor on the Recommendation for a Council Regulation amending Regulation (EC) No 2533/98 of 23 November 1998 concerning the collection of statistical information by the European Central Bank 2009/C 192/01.

does not go against the core idea of a democratic constitutional state. This is, against the social contract between the citizen and the State that demands citizens cooperate with the government by means of, inter alia, providing their personal details in order to safeguard general interests (such as security, tax collection, organisation of society, etc.) while preventing the State to pass such data on to the commercial sector. In addition, many specific legal doctrines may come into play, such as intellectual property law and the rights to privacy and data protection. That the latter rights are at stake is evident for a number of reasons. Many, if not all, government documents contain personal data, especially as the interpretation of the term 'personal data' has been increasingly widened.

As often happens when there is a complicated interplay between two legal instruments of the European Union, the EU legislator does not make a choice but leaves the exact relationship between these two instruments in the middle. For example, in the 2019 text, Article 1(4) holds: 'This Directive is without prejudice to Union and national law on the protection of personal data, in particular Regulation (EU) 2016/679 and Directive 2002/58/EC and the corresponding provisions of national law.' A strict reading would imply that the publication of public sector information for re-use may never restrict citizens' right to data protection. This raises numerous fundamental questions because, although the GDPR does not lay down an absolute prohibition on the use and even re-use of information, numerous restrictions do apply, such as, but not limited to:

- The purpose limitation principle, by which data may only be processed for the same purpose for which they were collected (Article 5(1)(b) GDPR). The potential exceptions to this principle, including consent and the processing for scientific or historical research, will seldom apply to the reuse of public sector information for commercial purposes.
- The data minimisation principle, which specifies that data may only be processed insofar as this is strictly necessary for the purpose for which they were collected (Article 5(1)(c) GDPR). The purposes for which data have been collected will vary from case to case but will typically concern matters such as taxation, providing social benefits, and protecting public order. Making available the data for reuse is usually not strictly necessary in light of these public interests.
- Moreover, the release of government information entails that there is no control over the purposes for which the information will be processed by third parties. Article 5(1)(f) GDPR states the principle of integrity and confidentiality, ensuring that unauthorised third parties cannot gain access to personal data. Publishing information online seems to run counter to this principle.
- Finally, the further re-use of the public sector information for a specific purpose, for example, the development of an app that can be downloaded in return for a small monthly payment and showing crime figures per city, district, and street, must have a legitimate processing basis. For the commercial re-use of 'ordinary' personal data, there will usually be only one ground that can be invoked, namely, the case referred to in Article 6 (1)(f) GDPR, where the interest in the re-use of the information for commercial purposes overrides the interests of the data subjects in the protection of their fundamental rights. A determination regarding the applicability of this provision will have to be made on a case-by-case basis, but it is clear that only in a limited number of cases can this ground be successfully invoked because the interests of citizens will weigh heavily, in particular, if data about children are being processed. In addition, the processing of 'special' or 'sensitive' personal data, such as those concerning race, religion, sexual orientation, or health is prohibited (Article 9 GDPR), and data relating to criminal convictions and offences can be only processed under the control of official authority or when the processing is authorised by Union or Member State law providing for appropriate safeguards for the rights and freedoms of data subjects (Article 10 GDPR). It is unclear whether any of the ten exception grounds mentioned in Article 9(2)GDPR will apply to most cases of reuse of public sector information for commercial purposes.

3.2.2.4 CJEU

Although there have been repeated calls to clarify the relationship between the two legal regimes, the EU legislator has chosen to leave the interpretation of this matter to national authorities and the courts. However, in the case of *Latvijas Republikas Saeima,*[102] the EU Court pronounced itself in relation to passive disclosures of data. It did so on the basis of preliminary questions from the Latvian court concerning the relationship between the regimes at the national level. The CJEU, in this case, questions the necessity of the processing of personal data and the subsidiarity of the regime adopted by the Latvian Parliament. When it comes to the protection or improvement of road safety, the Court states, that the legislation of other Member States shows that less intrusive measures than making available information on persons concerning their traffic offences may suffice. The Court emphasises that making this information public can lead to stigmatisation and other social consequences. It also questions the causal relationship between the regime established by the Latvian Parliament and the decline in traffic offences in Latvia. Next, it argues that the regime allows third parties to access the information even if those third parties have other purposes than those related to increasing road safety. This is not allowed because of the purpose limitation principle, the Court points out.

The Court also refers to two provisions that allow the Member States to restrict the right to data protection, namely in light of the freedom of expression (Article 85 GDPR) and of access to governmental information (Article 86 GDPR). Can the Latvian regulation be seen as staying within the discretionary power left to the Member States by those provisions? No, the Court of Justice rules. Whilst, as follows from recital 154 GDPR, public access to official documents constitutes a public interest capable of justifying the disclosure of personal data contained in such documents, that access must nevertheless be reconciled with the fundamental rights to respect for private life and to the protection of personal data, as Article 86 GDPR indeed expressly requires. In the light of the sensitivity of data relating to penalty points imposed for road traffic offences and of the seriousness of the interference with the fundamental rights of data subjects to respect for private life and to the protection of personal data, which is caused by the disclosure of such data, it must be held that those rights prevail over the public's interest in having access to official documents, in particular the national register of vehicles and their drivers. Furthermore, for the same reason, the right to freedom of information referred to in Article 85 GDPR cannot be interpreted as justifying the disclosure to any person who so requests personal data relating to penalty points imposed for road traffic offences. The Court, having stressed that giving citizens access to sensitive data concerning other citizens at their request, without them having to specify their interest, let alone prove that it is a legitimate interest, is not legitimate, points out that the same applies to the Latvian practice of passing on traffic safety information to commercial parties. The Directive on the re-use of public sector information, the Court once again emphasizes, indicates that the GDPR should be fully respected.

3.3 Technical developments

This section will provide insights gained on the technologies that can be used for anonymising and de-anonymising gained through the literature study (section 3.3.1), the interviews conducted for this study (section 3.3.2) and a workshop held for this study (section 3.3.3).

3.3.1 Literature study

This section will discuss two methods, namely data aggregation (section 3.3.1.1) and statistical disclosure control (section 3.3.1.2).

3.3.1.1 Data aggregation

[102] CJEU, C-439/19 - *Latvijas Republikas Saeima* [2021] ECLI:EU:C:2021:504.

Data aggregation can, in a way, be seen as a privacy-preserving method to use data without identifying individuals. Research on the privacy-preserving aggregation that guarantees privacy, confidentiality, security, and integrity has been carried out in the last decades, and different approaches and methods have been proposed.[103] The use of aggregate data aims to enable the processing of personal data at the highest level of abstraction and with the least possible detail in which it is still useful and privacy-preserving.[104] Two approaches have been proposed in the literature to the identifiability test, namely the *absolute*[105] and *relative*[106] approaches. Related to the formal notion of identifiability, privacy and data protection risks inherent to the categories of personal information have been postulated. Aggregate data presents the lowest privacy risks to the fundamental rights and freedoms of individuals, given the difficulty in singling out the individual and in inferring specific information about him or her. Therefore, aggregate data ostensibly occupies a prominent position as non-identifiable information.

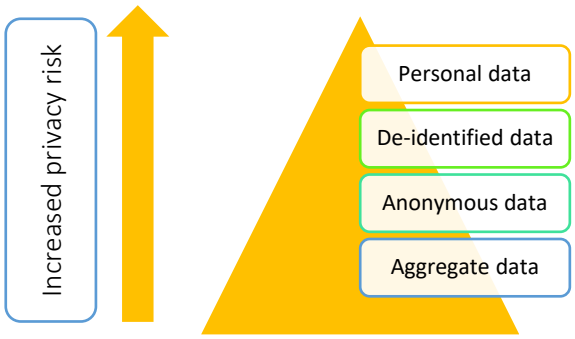

Figure 4: Increased Privacy Risks[107]

The original dataset from which aggregate data is to be created has to be prepared by performing various extractions and transformations to make it suitable for data mining. In this process, most of the issues relating to the creation and transformation of a dataset are related to summarization, aggregation, denormalization, and crosstabulation.[108] Four steps can be differentiated in the preparation of a dataset:

a.) data selection, where a selection of appropriate data is performed;
b.) data integration, where the collected data from different sources are combined and stored inside a table;
c.) data transformation, where data is transformed into the format required for each operation; and
d.) data reduction, where the data is compressed for the easiness of the analysis. Upon the completion of the preparation steps, data can be further computed or shared.

Ventura Silva et al. differentiate between three types of data aggregation approaches, *i.e.,* aggregation based on third parties, data perturbation, and cryptography.[109]

- In the first approach, trusted third parties may collect raw data, aggregate these data, and transfer the resulting data to authorized recipients. In this way, recipients only have the aggregate data. However, this might not be the case for a trusted third party. Although the use of pseudonyms has been proposed in this area,[110] reliance on a third trusted party remains a flaw in this model.

[103] Z. Erkin, J. R. Troncoso-pastoriza, R. L. Lagendijk and F. Perez-Gonzalez, Privacy-preserving data aggregation in smart metering systems: an overview, in IEEE Signal Processing Magazine, vol. 30
(2) 115.
[104] ENISA (2014) Privacy and Data Protection by Design – from policy to engineering, December, 20.
[105] The absolute approach is characterized by and abstraction from the concrete possibilities of identifiability given a processing operation to embrace any theoretical and/or remote probability of identifiability outside the scope of such processing operation. As such, all theoretical chances of combining data to identify the natural person are taken into account.
[106] The relative approach considers the particular circumstances of the processing operation as well as the probabilities of identifiability in relation to the concrete processing operation. As such, only realistic possibilities of combining data to identify the natural person are taken into account.
[107] Compiled by the authors based on Stallings, W. (2019). Information privacy engineering and privacy by design: Understanding privacy threats, technology, and regulations based on standards and best practices. Addison-Wesley Professional.
[108] Kuttappan, A. P., & Saranya, P. (2015). An Overview of various methodologies used in Data set Preparation for Data mining Analysis. International Research Journal of Engineering and Technology (IRJET), 2(2), 947-952, p. 948.
[109] Silva, L. V., Marinho, R., Vivas, J. L., & Brito, A. (2017). Security and privacy preserving data aggregation in cloud computing. In Proceedings of the Symposium on Applied Computing (pp. 1732-1738).
[110] Bohli, J. M., Sorge, C., & Ugus, O. (2010). A privacy model for smart metering. In 2010 IEEE International Conference on Communications Workshops (pp. 1-5). IEEE.

- In the second approach, random noise is added to the collected data so that the original data is not traceable, but aggregate values may still be calculated with a small or negligible error. The drawback of data perturbation is the difference between the original data and the perturbed data, which may lead, in certain cases, to disparities in the computation.
- In the third approach, cryptographic primitives can be used to overcome the drawbacks of the previous methods. Two approaches are common in this realm: secret sharing schemes and fully homomorphic encryption. Secret sharing schemes are based on centralized models, for example, where the aggregator is a third trusted party, personal data are completely hidden from the aggregator since it receives only encrypted data that it cannot decrypt and random shares of the total input. The drawbacks of this scheme are primarily scalability and communication overhead. On the other side, fully homomorphic encryption (FHE) is an encryption technology that allows the performance of an analysis '*in the ciphertext in the same way as in the plaintext without sharing the secret key*.'[111] This implies that the computation is performed over the encrypted data without the need to decrypt it, thus enabling data sharing with third parties. The results of the computation are equally encrypted so that only the exporters of data are able to decrypt the data. It is for this reason that some authors have acknowledged the processing of fully homomorphic encrypted data as falling out of the scope of the GDPR.[112] However, according to the state of the art, FHE is still highly inefficient and cannot be seen as a practical alternative to the processing of plaintext.[113] Only a limited number of computations on ciphertexts, such as polynomial operations, have reached sufficient precision for their use in practical scenarios.[114]

3.3.1.2. Statistical Disclosure Control

Statistical Disclosure Control (SDC) technologies aim at eliminating both directly and indirectly identifying information in a dataset while preserving data quality as much as possible.[115] Public institutions responsible for collecting and analysing statistical data have some form of a statistical disclosure control policy. In the Netherlands, Statistics Netherlands makes use of a Statistical Disclosure Control Handbook, which contains statistical disclosure control methods for microdata, quantitative tables, frequency tables, and analysis results.[116] According to Statistics Netherlands, statistical disclosure control entails preventing content-related conclusions about recognisable units are made based on published or otherwise available data of Statistics Netherlands.[117] The specialist in charge of protecting the data has to use different disclosure control methods in such a way that the minimum required level of protection is achieved and that the information loss is as small as possible, which will differ per situation. What constitutes information loss cannot be determined as such, as information is a subjective term which can be defined differently by each user. That makes it difficult to prescribe a specific method for a specific situation, rather, methods can be assessed based on their general advantages and disadvantages. For example, for microdata, it first has to be determined whether the disclosure is possible at all, which will depend on whether there are respondents that can be recognised as unique or rare cases in the microdata. Second, the variables that can potentially be used to identify a respondent have to be assessed. Combinations of categories of identifying variables can lead to unique or rare people. For example, 'mayor in Amsterdam' is unique, or 'female neurosurgeon older than 55 years of age from Staphorst' is considered rare. Rare combinations have to occur sufficiently often in the target population. By combining categories of identifying variables, rare combinations can be made

[111] ENISA (2015)Privacy by design in big data – An overview of privacy enhancing technologies in the era of big data analytics, 40. See also: Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. In Proceedings of the forty-first annual ACM symposium on Theory of computing, 169-178.
[112] Spindler, G., & Schmechel, P. (2016). Personal data and encryption in the European general data protection regulation. J. Intell. Prop. Info. Tech. & Elec. Com. L., 7, 163.
113 ENISA (2014) Privacy and Data Protection by Design – from policy to engineering, 43.
[114] Scheibner, J., Raisaro, J. L., Troncoso-Pastoriza, J. R., Ienca, M., Fellay, J., Vayena, E., & Hubaux, J. P. (2021). Revolutionizing medical data sharing using advanced privacy-enhancing technologies: technical, legal, and ethical synthesis. Journal of medical Internet research, 23(2), e25120.
[115] Bargh, M.S., Meijer, R., Vink, M (2018). On statistical disclosure control technologies: For enabling personal data protection in open data settings. WODC Cahiers,20.
[116] Hundepool, A., Jonker, J., Nobel, J., Schulte Nordholt E. and De Wolf, P.P. (2006). Handboek Statistische Beveiliging, Statistics Netherlands, Voorburg.
[117] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & De Wolf, P. P. (2012). Statistical disclosure control (2). New York: Wiley.

less rare. There are several techniques that can be used to prevent identification in microdata, such as global recording, local suppression, top-coding, adding noise to weights, or using the Post Randomisation Method.

In their work on different SDC methods, Domingo-Ferrer and Torra propose to distinguish between two types of statistical output to be protected with SDC techniques: microdata and tabular data. To assess methods for statistical data protection, basic attributes are important, the disclosure risk and utility. 'Disclosure risk: A measure of the risk to respondent confidentiality that the data releaser (typically a statistical agency) would experience as a consequence of releasing the table. Data utility: A measure of the value of the released table to a legitimate data user.'[118] For good microdata protection, they propose empirical disclosure risk measures based on record linkage instead of relying on measures that are dependent on uniqueness. For tabular data protection, they demonstrated that the most widely used sensitivity rule for a priori risk assessment, the dominance rule, is flawed, and other stronger measures are necessary.

In addition to protecting microdata, quantitative tables, and frequency tables, there is still a very broad, diverse group of statistical output to be protected. These results also run a risk of disclosing the data for individual respondents and are treated with care, but, in practice, there are always two risks: on the one hand, the publishing of risky results could incorrectly be approved, or safe results could incorrectly be held back. The balance between disclosure risk and data utility is an important but difficult one to draw.[119] Thus, the use of SDC in combination with different attacker scenarios should prevent the disclosure of unintended information.

3.3.2 Results from interviews

This section will discuss some of the main findings gained through the interviews conducted for this study. The full interview reports may be found in the annex to this report.

Though anonymous (micro) data and aggregate (macro) data are clearly different in technical terms, legal regulation treats them the same. Yet there are different risks attached to disclosing micro and macrodata. Some experts suggested developing more concrete rules for disclosing aggregated data, such as having a minimum number of people in a cell with a frequency count table or rules on dominance with quantitative magnitude tables. In addition, checks for group disclosure could be stipulated.

A major challenge for NSIs is the increased availability of open data available, which makes it hard to assess the availability of data when realising aggregated datasets. The most important tool for safeguarding information in statistical data is statistical disclosure control (SDC). Public use files are intended for the public at large, so for informational or educational purposes. Within those, microdata files have to be very protected in terms of statistical disclosure control. There should be no, or virtually no, possibilities of identifying persons in those data, and certainly, there should be no sensitive information in that dataset. There are also tabular data on Statline, which is open data in some sense. Those data are protected against disclosure even if combined with other information. The level of SDC that is applied depends on the level of legal protection that is provided: the more legal protection, the less SDC is necessary. Disclosure is usually linked to an attacker scenario. There are still discussions about whether differential privacy can be used in official statistics because of the loss of utility of the data, while accuracy has to be high.

[118] Domingo-Ferrer, J., & Torra, V. (2004). Disclosure risk assessment in statistical data protection. Journal of Computational and Applied Mathematics, 164, 285-293.
[119] George T. Duncan & S. Lynne Stokes (2004). Disclosure Risk vs. Data Utility: The R-U Confidentiality Map as Applied to Topcoding, CHANCE, 17:3, 16-20.

3.3.3 Results from the workshop

The workshop held for this study yielded the following results:

- Complexity of the terminology: the similar regulation of anonymous and aggregate data raised questions among technical experts as well as the exact boundary between personal or identifiable data and aggregated or statistical data.
- Conflicting messages: the legal regime gives conflicting signals. On the one hand, open data, re-use of public sector information, and data portability are promoted, and, on the other hand, privacy, secrecy, and data protection are emphasised. The regulator should provide more clarity as to which choices should or should not be made in practice.

3.4 Analysis

Five main tensions between the legal and the technical realm have emerged from this chapter:

1. Privacy and data protection regimes have traditionally focused on natural persons and identifiability. This choice is understandable given that, traditionally, there were, by and large, two types of data processing. Individual, specific data collection, for example, by law enforcement agencies on suspects or by companies on their customers, and statistical data processes, performed inter alia for developing models, maps, and fact-based governmental policies. It was the first type of data processing that privacy and data protection regimes focussed on. Since then, however, at least two things have changed. First, it is increasingly easy to infer specific individual information and sometimes sensitive personal data about natural persons from aggregate data, especially when combining a dataset with aggregate data with another data source. Second, data practice has moved from collecting individual data to focussing on aggregated data, group profiles, and longitudinal patterns. These are used for increasingly intensive and far-reaching decisions that affect people as part of a group or category.
2. In addition, with new developments in technology, it might become increasingly possible to reidentify individuals in aggregated data.
3. The same holds true for the fact that through evolving technological capacities, it is increasingly possible to arrive at personal data by combining two or more datasets that, in isolation, do not contain any personal data.
4. Statistics are used to generate knowledge by analysing existing data to make assumptions about individuals, for example, by mapping past experiences and establishing correlations between certain characteristics and particular outcomes or behaviour.[120] AI and big data analytics allow people to be profiled in actionable ways without being personally or individually identified.[121] This means that even aggregated data that are not re-identified can be qualified as falling under the data protection framework. In essence, this development, as well as those discussed under points 2 and 3, especially when seen in the light of each other, means that more and more, aggregated or statistical data should be deemed to fall under the data protection framework per sé, even if no identifying information is contained in it.
5. Finally, what makes these tensions more complex is that legislators and courts do not present a uniform view as to what extent collection and use of aggregate data should be regulated or to what extent aggregate data can also be personal data.[122] For example, the European legislator leaves room for the Member States here to determine safeguards, and this space is used differently by different national legislators and courts. An important question to answer is what

[120] European Union Agency for Fundamental Rights, Preventing unlawful profiling today and in the future: a guide (2018).
[121] See for example: Strandburg, K. (2014). Monitoring, datafication and consent: legal approaches to privacy in the big data context. In Lane, J., Stodden, V., Bender, S., Nissenbaum, H. (Eds.). (2014). Privacy, Big Data, and the Public Good: Frameworks for Engagement. Cambridge University Press. Barocas, S., Nissenbaum, H. (2014) Big data's end run around anonymity and consent. In Lane, J., Stodden, V., Bender, S., Nissenbaum, H. (Eds.). (2014) Privacy, Big Data, and the Public Good: Frameworks for Engagement. Cambridge University Press;
[122] Finck, M., & Pallas, F. (2020). They who must not be identified—distinguishing personal from non-personal data under the GDPR. International Data Privacy Law. Stalla-Bourdillon, S., & Rossi, A. (2021). Aggregation, synthesization and anonymization: a call fora risk-based assessment of anonymization approaches.

the role of statistical or aggregated data is in society, as this determines which techniques should be used to analyse or protect such information and if the current level of legal safeguards in various jurisdictions and instruments is sufficient in relation to the value and possible impact of such data. It is clear that aggregate and statistical data have value from many perspectives and for different actors, such as an informative value to the public, a starting point for developing policies and decision-making by public actors or specific interests to private parties, such as using statistics in support of a legal claim. However, statistical data, especially when made public, entails a risk for re-identification and data composition.[123]

[123] Duncan, G. T., Jabine, T. B., & de Wolf, V. A. (Eds.). (1993). Private lives and public policies: Confidentiality and accessibility of government statistics. National Academy Press.

Chapter 4: Pseudonymization and de-pseudonymization

4.1 Introduction

This chapter describes the boundary between pseudonymised data and non-personal data, and between pseudonymised data and personal data. The GDPR is applicable to pseudonymised data, but some exceptions apply. In addition, pseudonymisation is regarded as one way to implement technical organisational security standards within an organisation. This chapter will discuss technologies for pseudonymisation and depseudonymisation, the way in which pseudonymous data are currently regulated, the gaps that exist between the regulatory framework and the technical reality, and potential solutions to solve that gap, section 4.2 discusses the legal distinction between pseudonymous and non-pseudonymous data, section 4.3 describes the main techniques available for pseudonymising and de-pseudonymising data and section 4.4 analyses the gap between the legal regulation and technical reality.

4.2 Legal regulation

This section will give an overview of the legislative history of European data protection law as far as relevant for pseudonymous data. It will start with a discussion of pre-GDPR regulation (section 4.2.1), reflections by the Article 29 Working Party (section 4.2.2), and the approach taken with respect to data breaches (section 4.2.3). Subsequently, it will discuss how pseudonymous data are regulated under the GDPR (section 4.2.4) and how an Advocate General of the CJEU has reflected on this notion (section 4.2.5). Finally, a reference is made to a report by ENISA (section 4.2.6).

4.2.1 Pre-GDPR laws

In legal terms, pseudonymisation could be best understood as a technical security measure. When unauthorised personnel or third parties get access to the data, they may need to invest considerably to infer the identity of the data subject. In this sense, it is similar to encryption, which means that the data controller has the key to decrypt the data, just like it has the identifier behind the pseudonym, but it ensures that unauthorised parties cannot or will have difficulty deciphering the dataset. Such security standards, both in organisational and technical sense, to disable unauthorised personnel and third parties from having access to the data and to ensure that, when they do nevertheless, they have difficulty to act on them, have been part of the data protection frameworks ever since the 1970ties and the Convention from 1981. But it was only under the EU frameworks that these elements gained a special status. The proposal for the EU Directive underlined the fact that current computer-based techniques can offer a substantially higher degree of data security for specific individual requirements, such as sophisticated encryption techniques. The very first draft of the predecessor of the e-Privacy Directive also contained an obligation for providers to offer users end-to-end encryption. In the same package of proposals was a Council Decision on information security which was ultimately adopted in 1992, and , which advocated for the adoption of encryption techniques.[124]

One of the first uses of the notion of pseudonymous data was in a communication of the Commission to the Parliament on Privacy Enhancing Technologies in 2007: 'A further step to pursue the aim of the legal framework, whose objective is to minimise the processing of personal data and using anonymous or pseudonymous data where possible, could be supported by measures called Privacy Enhancing Technologies or PETs - that would facilitate ensuring that breaches of the data protection rules and violations of individual's rights are not only something forbidden and subject to sanctions, but

[124] COM(90) 314 final ~.sYN 287 and 288 Brussels, 13 September 1990. 92/242/EEC: Council Decision of 31 March 1992 in the field of security of information systems.

technically more difficult.'[125] This was referred to in a number of other legislative documents, such as the Commission Recommendation on the implementation of privacy and data protection principles in applications supported by radiofrequency identification, which in a recital stressed the 'goal of minimising the processing of personal data and using anonymous or pseudonymous data wherever possible by supporting the development of PETs and their use by data controllers and individuals.'[126]

4.2.2 Article 29 Working Party

In its opinion on personal data, Article 29 Working Party distinguished between pseudonymous data and key-coded data. It is interesting that the Party still had a very limited understanding of pseudonymous data. It found that the aim of pseudonymization 'is to be able to collect additional data relating to the same individual without having to know his identity. This is particularly relevant in the context of research and statistics. Pseudonymisation can be done in a retraceable way by using correspondence lists for identities and their pseudonyms or by using two-way cryptography algorithms for pseudonymisation. Disguising identities can also be done in a way that no reidentification is possible, e.g., by one-way cryptography, which creates in general anonymised data.'[127] It qualified pseudonymised data as indirectly identifiable data. It treated key-coded data as a classic example of pseudonymisation, which it found was primarily used in the medical sector.

In its opinion on anonymisation techniques, the Article 29 Working Party explicitly stressed that pseudonymous data should not be considered equivalent to anonymised data. It referred inter alia to the America Online (AOL) incident. But it did support the increased use of pseudonymisation techniques, of which it distinguished between five important ones, including encryption, thus treating (certain forms of) encryption as a subset of pseudonymisation techniques:[128]

1. Encryption with secret key: the holder of the key can trivially re-identify each data subject through decryption of the dataset, but it may make such impossible or difficult for third parties.
2. Hash function: a function which returns a fixed size output from an input of any size (the input may be a single attribute or a set of attributes) and cannot be reversed; this means that the reversal risk seen with encryption no longer exists. However, the WP29 underlined, if the range of input values the hash function is known they can be replayed through the hash function in order to derive the correct value for a particular record. The use of a salted-hash function (where a random value, known as the 'salt', is added to the attribute being hashed) can reduce the likelihood of deriving the input value but nevertheless, but not impossible.
3. Keyed-hash function with stored key: a particular hash function which uses a secret key as an additional input (this differs from a salted hash function as the salt is commonly not secret). A data controller can replay the function on the attribute using the secret key, but it is much more difficult for an attacker to replay the function without knowing the key as the number of possibilities to be tested is sufficiently large as to be impractical.
4. Deterministic encryption or keyed-hash function with deletion of the key: selecting a random number as a pseudonym for each attribute in the database and then deleting the correspondence table. It will be computationally hard for an attacker to decrypt or replay the function, as it would imply testing every possible key.
5. Tokenization: this technique is typically applied in the financial sector to replace card ID numbers with values that have reduced usefulness for an attacker. It is derived from the previous ones being typically based on the application of one-way encryption mechanisms or the assignment, through an index function, of a sequence number or a randomly generated number that is not mathematically derived from the original data.

[125] **Communication from the Commission to the European Parliament and the Council on Promoting Data Protection by Privacy Enhancing Technologies (PETs) /* COM/2007/0228 final.**
[126] Commission Recommendation of 12 May 2009 on the implementation of privacy and data protection principles in applications supported by radiofrequency identification (notified under document number C(2009) 3200) (2009/387/EC).
[127] Working Party 29, 'Opinion 4/2007 on the concept of personal data', 20 June 2007.
[128] Working Party 29, 'Opinion 05/2014 on Anonymisation Techniques', 10 April 2014.

4.2.3 Data Breaches

In 2013, the Commission adopted a regulation on the measures applicable to the notification of personal data breaches under the e-Privacy Directive.[129] In it, there was an exemption from informing data subjects of a data breach when the data were made unintelligible. 'Data shall be considered unintelligible if: (a) it has been securely encrypted with a standardised algorithm, the key used to decrypt the data has not been compromised in any security breach, and the key used to decrypt the data has been generated so that it cannot be ascertained by available technological means by any person who is not authorised to access the key; or (b) it has been replaced by its hashed value calculated with a standardised cryptographic keyed hash function, the key used to hash the data has not been compromised in any security breach, and the key used to hash the data has been generated in a way that it cannot be ascertained by available technological means by any person who is not authorised to access the key.'[130]

4.2.4 Legislative process of the GDPR

In the legislative process of the GDPR, the notion of pseudonymous data was first discussed in the impact assessment by the Commission, in which reference was made to the legal status of pseudonymised data in the various Member States. It referred to a number of countries which treated these data only as personal data in relation to the data controller, the person or organisation with the key, and to a number of other countries where such data were treated as personal data per se, even if the data are processed by someone who has no means for such re-identification. 'However, DPAs in those Member States are generally less demanding with regard to the processing of data that are not immediately identifiable, taking into account the likelihood of the data subject being identified as well as the nature of the data.'[131] It was not the Commission but the Parliament that proposed to give the notion of pseudonymous data a more central role. Not only did it provide a definition, it also suggested including a reference to pseudonymous data in a recital and no less than five articles in the GDPR. This unleashed a fierce discussion in parliament itself, with dozens of amendments on the definition of pseudonymisation and pseudonymous data and its role in the various articles in the GDPR. The Council affirmed the position of pseudonymous data in the GDPR and suggested, in many places, to equate it with encrypted data[132] (which some members of Parliament suggested should also beget its own definition).[133]

In the final version of the GDPR, the notions of encryption and pseudonymization are indeed equated, at least in two instances. Paragraph 4 of Article 6 concerns the further processing of personal data. Such will be deemed legitimate with the consent of the data subject(s) or when there is an explicit legal basis, or when the original purpose and the new purpose for which the data were gathered can be deemed compatible.[134] In order to determine that, the controller should take account, inter alia, of 'the existence of appropriate safeguards, which may include encryption or pseudonymisation'. Article 32 GDPR, regarding the security of the data processing, suggests that the data controller should adopt adequate safety measures, such as 'the pseudonymisation and encryption of personal data'. Article 34 concerns the case in which a data breach has occurred. In principle, such data breach must be reported to the data

[129] Commission Regulation (EU) No 611/2013 of 24 June 2013 on the measures applicable to the notification of personal data breaches under Directive 2002/58/EC of the European Parliament and of the Council on privacy and electronic communications OJ L 173, 26.6.2013.
[130] Article 4 No 611/2013. See further: Working Party 29, 'Opinion 03/2014 on Personal Data Breach Notification', 25 March 2014.
[131] Brussels, 25.1.2012 SEC(2012) 72 final.
[132] Brussels, 8 April 2016 (OR. en) 5419/1/16 REV 1.
[133] 2012/0011(COD) 04.03.2013 AMENDMENTS (2) 602 – 885 Draft report Jan Philipp Albrecht (PE501.927v04-00). RR\1010934EN.doc PE501.927v05-00 EN United in diversity EN P7_TA(2014)0212
Protection of individuals with regard to the processing of personal data ***I European Parliament legislative resolution of 12 March 2014 on the proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) (COM(2012)0011 – C7-0025/2012 – 2012/0011(COD)).
[134] See also: Working Party 29, 'Opinion 03/2013 on purpose limitation', 2 April 2013.

subjects concerned unless the data are made unintelligible to third parties. The GDPR refers to encryption by way of example, but it seems likely that certain types of pseudonymisation would be treated similarly,[135] though recital 85 explicitly warns of the reversal of pseudonymisation as a risk of data breaches.

Article 25, concerning data protection by design and by default, mentions pseudonymisation as an example, though again, it seems that encryption could also have been included in the list. This also is the case with respect to Article 40, concerning the codes of conduct that sectors can adopt, in which it is suggested that such codes may lay down rules for the pseudonymisation of data, and with respect to Article 89, concerning the further processing of data for statistical, historical and archival purposes, for which technical and organisational security measures should be implemented, such as the pseudonymisation of data.[136]

In Article 4, the GDPR does not give a definition of encrypted data, but it does provide one for pseudonymisation, thus focusing on the process instead of the status of the data. Pseudonymisation means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person. Recitals 26, 28, and 29 provide further guidance on pseudonymisation. Importantly, with respect to the different choices in the various Member States with respect to the status of pseudonymised data, as mapped out in the impact assessment, the GDPR makes an explicit choice to treat such data as personal data.[137] Consequently, in principle, the GDPR applies in full, but recital 28 makes clear that pseudonymisation can reduce the risks to the data subjects concerned and help controllers and processors to meet their data-protection obligations.

4.2.5 CJEU

In *Tele2 Sverige AB*, Advocate General Saugmandsgaard Øe seemed sceptical of the approach in which certain safeguards taken may mean that other safeguards do not have to be taken, which he coined the 'communicating vessels argument'. 'The pernicious effect of the 'communicating vessels' argument may be easily illustrated by the following examples. A national regime that rigorously restricts access to the service of the fight against terrorism and limits the retention period to three months (representing a strict approach to access and retention period), but does not require service providers to retain the data, in encrypted form, within the national territory (representing a flexible approach to security), would expose the entire population to a significant risk of the retained data being accessed unlawfully. Similarly, a national regime that provided for a retention period of three months and the retention of the data in encrypted form within the national territory (representing a strict approach to retention period and security), but which allowed all employees of all public authorities access to the retained data (representing a flexible approach to access), would expose the entire population to a significant risk of abuse on the part of the national authorities.'[138]

4.2.6 European Union Agency for Cybersecurity (ENISA)

[135] See for further guidance: EDPS Guidelines on personal data breach notification For the European Union Institutions and Bodies, 21 November 2018. EDPS, Guidelines on the protection of personal data in IT governance and IT management of EU institutions, 23 March 2018.
[136] Compare Article 4 para 1 sub e Regulation (EC) No 45/2001 of the European Parliament and of the Council of 18 December 2000 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data.
[137] See further: CJEU, C-673/17, *Bundesverband der Verbraucherzentralen und Verbraucherverbände - Verbraucherzentrale Bundesverband e.V. v Planet49 GmbH* [2019] ECLI:EU:C:2019:246; CJEU, C-520/18, Ordre des barreaux francophones and germanophone and Others. Opinion of Advocate General [2020] ECLI:EU:C:2020:7. CJEU, Case C-215/20: Request for a preliminary ruling from the Verwaltungsgericht Wiesbaden (Germany) lodged on 19 May 2020 — JV v Bundesrepublik Deutschland.
[138] CJEU, C-572/14, Austro-Mechana Gesellschaft zur Wahrnehmung mechanisch-musikalischer Urheberrechte Gesellschaft mbH v Amazon EU Sàrl and Others [2016] ECLI:EU:C:2016:572.

A report on pseudonymization techniques and best practices was issued in 2019 by ENISA. This report contains recommendations on shaping technology according to data protection and privacy provisions. One of the main conclusions from the report is that it requires a high level of competence in order to apply a robust pseudonymization process, possibly reducing the threat of discrimination or re-identification attacks while maintaining the degree of utility necessary for the processing of the pseudonymized data. ENISA vouches for a risk-based approach to pseudonymization: 'Data controllers and processors should carefully consider the implementation of pseudonymization following a risk-based approach, taking into account the purpose and overall context of the personal data processing, as well as the utility and scalability levels they wish to achieve.'[139] It also gives a number of examples of pseudonymization scenarios. Additionally, it touches upon the main types of attacks that can be done with a pseudonymisation technique. By providing these examples, it gives one a better idea of what to pay attention to when picking a pseudonymization technique.

4.3 Technical developments

This section will provide insights gained on the technologies that can be used for pseudonymising and de-pseudonymising and encryption and decryption gained through the literature study (section 4.3.1), the interviews conducted for this study (section 4.3.2) and a workshop held for this study (section 4.3.3).

4.3.1 Literature study

This section will discuss general techniques for pseudonymisation and encryption (section 4.3.1.1) and de-encryption, especially in light of quantum computing (section 4.3.1.2).

4.3.1.1 Pseudonymization techniques

While pseudonymisation aims at replacing personal identifiers with pseudonyms to decrease linkability among information, encryption focuses on rendering personal data unintelligible to prevent its access. As a result, despite the lack of attribution between pseudonyms and initial identifiers, pseudonymization may not prevent third parties from identifying the data subject with additional information nor conducting data mining operations over pseudonyms in order to extract information about the non-identified data subject.[140] Encryption, on the other hand, focuses on the confidentiality of personal data. It is intended to prevent the disclosure of personal data to unauthorized parties and, thus, the possibility of accessing and using it. This implies that, while pseudonymisation aims to hide the identity of the natural person, encryption aims at hiding the whole dataset. Consequently, encryption may be used as a pseudonymisation technique, whereas the opposite is not possible.

Pseudonymisation techniques can be classified along the following lines.

4.3.1.1.1 Hashing

4.3.1.1.1.1 Hash function

Hashing is a technique that can be used to derive pseudonyms. In a nutshell, hash functions are functions that compress an input of arbitrary length to a result with a fixed length. This fixed-size output is called message digest, hash value, hash code, or simply hash. In this way, if an identifier *m* is used as an input in the hash function *h,* the function will return a fixed-size pseudonym *h(m).* It is important to note that, for any given hash function, inputs which are the same result in the same hashes. Therefore, if the initial

[139] ENISA, 'Pseudonymisation techniques and best practices' Recommendations on shaping technology according to data protection and privacy provisions, 2019. <https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices>.
[140] ENISA, Recommendations on shaping technology according to GDPR provisions An overview on data pseudonymization, November 2018. ENISA, Data Pseudonymisation: Advanced Techniques & Use Cases. Technical analysis of cybersecurity measures in data protection and privacy. January 2021.

identifier *m* represents, for instance, the height of individuals in a given population, any individual sharing the same height would have the same hash value *h(m)*. This is better illustrated in the following Table 6, where individuals x_1, x_3, and x_5 share the same hash value $h(m(x_1))=h(m(x_3))=h(m(x_5))$ = 7b69759630f869f2723875f873935fed29d2d12b10ef763c1c33b8e0004cb405.

Individual	Initial identifier *m* Height (cm)	Hash function *h*	Hash *h(m)* - Output hash size 256
x_1	180	SHA-2	7b69759630f869f2723875f873935fed29d2d12b10ef763c1c33b8e0004cb405
x_2	179	SHA-2	3068430da9e4b7a674184035643d9e19af3dc7483e31cc03b35f75268401df77
x_3	180	SHA-2	7b69759630f869f2723875f873935fed29d2d12b10ef763c1c33b8e0004cb405
x_4	181	SHA-2	017242aed0751adb88388d165183d00ebec8345e029638bf0d0688afdbf91deb
x_5	180	SHA-2	7b69759630f869f2723875f873935fed29d2d12b10ef763c1c33b8e0004cb405
x_6	178	SHA-2	2093474895a9cef09980364d47d6a01723022d4a6617503302ea3f24274eb339

Table 6: Operation of hash function with SHA-2

ENISA suggests that hash functions with known vulnerabilities, such as MD5 and SHA-1 should be avoided[141] and, instead, replaced by cryptographically resistant hash functions, such as SHA-2 and SHA-3.[142]

The validity of hashing as a recommendable pseudonymisation technique is a controversial issue. In most cases, hashing is seen as a weak pseudonymisation technique. In this respect, Demir et al. have established the pitfalls of hashing by arguing three reasons for failure.[143]

- First, they contend that the properties of hash functions are commonly misunderstood. One of the assumptions of hashing is that the mathematical function supporting the generation of the hash or pseudonym is irreversible. This property is called one-wayness, and implies that the conversion of the original identifier into the pseudonym cannot, in theory, be inverted. In other words, it is computationally infeasible to generate the original identifier from the hash or pseudonym, *e.g.* in Table 6, a third party introducing *h(m)*= 7b69759630f869f2723875f873935fed29d2d12b10ef763c1c33b8e0004cb405 in the hash function *h*=SHA-2 would not be able to obtain the height *m*=180 cm. According to the authors, data custodians often underestimate the risk of exhaustive searches, which can render one-wayness reversible.
- Second, the authors claim that, even when exhaustive searches cannot be carried out on the initial domain space, it should however, be possible to do so on one of its subdomains. To this extent, an adversary determining whether a pseudonym belongs to a certain group or another may learn a discriminatory property about it based on its associated subdomain. This type of attack is better known as a discrimination attack, and it assumes that a certain domain *A* is split into two subdomains A_1 and A_2. If an identifier verifies a certain discriminating property, it will belong to A_1, whereas, otherwise, it will fit in A_2. In this respect, an illustrative scenario of this

[141] Wang, X., & Yu, H. (2005, May). How to break MD5 and other hash functions. In Annual international conference on the theory and applications of cryptographic techniques (pp. 19-35). Springer, Berlin, Heidelberg. See also Stevens, M., Bursztein, E., Karpman, P., Albertini, A., & Markov, Y. (2017, August). The first collision for full SHA-1. In Annual international cryptology conference (pp. 570-596). Springer, Cham.
[142] ENISA, Recommendations on shaping technology according to GDPR provisions An overview on data pseudonymization, November 2018, p. 21.
[143] Demir, L., Kumar, A., Cunche, M., & Lauradoux, C. (2017). The pitfalls of hashing for privacy. IEEE Communications Surveys & Tutorials, 20(1), 551-565.

type of attack is to assume that the adversary holds a list of identifiers which can be checked against the pseudonyms to see if they have been included in the dataset.

- Third, the authors are of the opinion that hashing cannot take into account any prior adversary knowledge, therefore undermining the adversarial strength.

Another important points to consider while evaluating any pseudonymisation and encryption technique are, according to ENISA, whether (i) third parties can reproduce the pseudonyms that a data controller creates across domains and (ii) whether the pseudonyms used can be easily re-identified.[144]

Hashing doesn't normally meet those standards. The difficulty in reproducing pseudonyms plays an important role in ensuring security, as it prevents third parties that apply the same hash function to use the generated pseudonyms across domains. This is better illustrated in the so-called brute force and dictionary attacks,[145] where an attacker using the same hash function as the controller or processor tries to introduce a large number of likely possibilities until the generated pseudonym is matched. By referring again to Table 6, a potential attacker applying the same *h*=SHA-2 hash could systematically enter each height until the value 7b69759630f869f2723875f873935fed29d2d12b10ef763c1c33b8e0004cb405 is matched. Since any third party that applies the same hash function to the same identifier gets the same pseudonym, the first property is not satisfied by hashing. In relation to the second property, it is also unlikely to hold since it would also be trivial for any third party to discover the correspondence between a given identifier, *e.g.* 180 cm, and its resulting pseudonym by simply hashing the identifier with the *h*=SHA-2 hash function. Based on these outcomes, hash functions are generally not recommended for pseudonymisation of personal data, without prejudice to their value as security-enhancing techniques in specific contexts with negligible privacy risks.[146]

4.3.1.1.1.2 Hashing with key or keyed hashing

Hashing with key or keyed hashing builds on the conventional hashing introduced above by adding a secret key that alters the output of the function *h*. On this basis, hashing with key can produce different pseudonyms for the same input according to the choice of the specific key. To illustrate this, see Table 7 below, where for the same initial identifiers m=180 cm of individuals x_1, x_3, x_5, the resulting $h(k,m)$ are different in all cases, according to the secret keys k used, namely $k_1, k_3,$ and k_5.

Individual	Initial identifier *m* Height (cm)	Secret Key *k*	Hash function *h*	Keyed Hash *h(k, m)* - Output hash size 256
x_1	180	ROT131HAV	SHA-2	8e612bd1f5d132a339575b8dafb7842c64614e56bcf3d5ab65a0bc4b34329407
x_2	179	PASS45BEE	SHA-2	620c9c332101a5bae955c66ae72268fbcd3972766179522c8deede6a249addb7
x_3	180	LONE23COM	SHA-2	ed0f61e6f6796d3d9f1ec1eb3851c8743e8b78c793741b0b4ba541e9e8a0313c
x_4	181	RABBIT32F	SHA-2	210e3b160c355818509425b9d9e9fd3ea2e287f2c43a13e5be8817140db0b9e6
x_5	180	THINK3MEN	SHA-2	87574c1abffa14d93d932b1f75f4360b83c6d1ccf3e514c6ca4de4081a9fbd31

[144] ENISA, Recommendations on shaping technology according to GDPR provisions An overview on data pseudonymization, November 2018, 19.
[145] ENISA, Data Pseudonymisation: Advanced Techniques & Use Cases. Technical analysis of cybersecurity measures in data protection and privacy. January 2021, 13.
[146] ENISA, Recommendations on shaping technology according to GDPR provisions An overview on data pseudonymization, November 2018, 22.

x_6	178	BIG249ALM	SHA-2	0fecf9247f3ddc84db8a804fa3065c013baf6b7c2458c2ba2bf56c2e1d42ddd4

Table 7: Operation of keyed hash function with SHA-2

The choice of the secret key cannot be trivial; it needs to be unpredictable and of sufficient length.[147] According to technical scholarship, secret keys of more than 256 bits of length are to be considered secure even in the post-quantum era.[148] Security of the secret key should be ensured to avoid keyed hashing being rendered conventional hashing. Hashing with key is a pseudonymization technique offering several advantages over conventional hashing. On the one side, the reproducibility of the pseudonym by third parties is prevented as the generated pseudonyms are different, therefore ensuring unlinkability across domains. On the other side, any third party without the knowledge of the key would not be in a position to reveal the original identifier from the pseudonym. For these reasons, hashing with key is generally considered a robust pseudonymisation technique from a data protection point of view.[149]

One of the most outstanding properties of keyed hashing that may place it in the crosshairs of anonymization techniques is its robust computational security. For instance, if the secret key is securely destroyed and the hash function is cryptographically strong, it would be computationally hard, even for the data controller, to reverse the pseudonym to the initial identifier.[150] This may also be the case even where the controller has knowledge of the initial identifiers. As a result, the use of keyed hashing together with the deletion of the secret key or salt may be considered anonymisation due to the rupture of the link between the pseudonym and the initial identifier.

4.3.1.1.1.3 Hashing with salt

Hashing with salt is a pseudonymisation technique that is a variant of keyed hashing, where a conventional hash function together with a so-called 'salt', or auxiliary random-looking data, is used. Just like keyed hashing, hashing with salt produces several pseudonyms for the same initial identifier. Therefore, hashing with salt enjoys the same properties as keyed hashing as long as the salt is appropriately secured and third parties do not have knowledge of it. Consider Table 8 below for illustrative purposes, where the same initial identifiers m=180 cm of individuals x_1, x_3, x_5 result in different $h(s,m)$ in all cases, according to the salt s used, namely s_1, s_3, and s_5.

Individual	Initial identifier m Height (cm)	Salt s	Hash function h	Salted Hash $h(s, m)$ - Output hash size 256
x_1	180	f1nd1ngn3m1	SHA-2	89aa1e580023722db67646e8149eb246c748e180e34a1cf679ab0b41a416d904
x_2	179	ab43bj36k34	SHA-2	1be00341082e25c4e251ca6713e767f7131a2823b0052caf9c9b006ec512f6cb
x_3	180	u96kv96hk99	SHA-2	a665a45920422f9d417e4867efdc4fb8a04a1f3fff1fa07e998e86f7f7a27ae3
x_4	181	r345l345cc4l	SHA-2	6affdae3b3c1aa6aa7689e9b6a7b3225a636aa1ac0025f490cca1285ceaf1487
x_5	180	krbkuu63233	SHA-2	a5e45837a2959db847f7e67a915d0ecaddd47f943af2af5fa6453be497faabca

[147] ENISA, Recommendations on shaping technology according to GDPR provisions An overview on data pseudonymization, November 2018, 23.
[148] Mavroeidis, V., Vishi, K., Zych, M. D., & Jøsang, A. (2018). The impact of quantum computing on present cryptography. arXiv preprint arXiv:1804.00200. See also NIST Special Publication (SP) 800-57 Part 1 Revision 4, Recommendation for Key Management – Part 1: General, National Institute of Standards and Technology, Gaithersburg, Maryland, January 2016, 160.
[149] ENISA, Data Pseudonymisation: Advanced Techniques & Use Cases. Technical analysis of cybersecurity measures in data protection and privacy. January 2021, 13.
[150] ENISA, Recommendations on shaping technology according to GDPR provisions An overview on data pseudonymization, November 2018, 24.

x_6	178	23vjd3od odo	SHA -2	65a699905c02619370bcf9207f5a477c3d67130ca71ec6f7 50e07fe8d510b084

Table 8: Operation of salted hash function with SHA-2

In addition to the share properties of keyed and salted hashing, two main drawbacks of salted hashing should be stressed. First, it is considered that keyed hash functions are more robust from a cryptographical point of view than salted hash functions. Although cryptographical techniques are available that generate strongly salted hashes, the fact that the salt does not share the same unpredictability properties as secret keys remains an important factor. Second, ENISA is the opinion that, in most common scenarios, salts are stored together with pseudonyms. This results in a serious threat to confidentiality. It is, therefore, suggested that salted hashes are used with carefulness and in accordance with available best practices.

4.3.1.1.1.4 Hashing with pepper

An alternative methods to salted hashing is peppered hashing. This pseudonymisation technique consists of adding a secret to the salt during the hashing and storing it separately from salts and pseudonyms in another medium, for instance, in a hardware security module. The pepper, therefore, shares certain properties with salt in that it is a random value and is similar to an encryption key in that it must be kept secret. Consider Table 9 below in relation to peppered hashing. Note that the pepper *p* should always be stored separately.

Individual	Initial identifier *m* Height (cm)	Pepper *p*	Salt *s*	Hash function *h*	Peppered Hash *h(p, s, m)* - Output hash size 256
x_1	180	flk34s nio1	f1nd1ng n3m1	SH A-2	eeca91fd439b6d5e827e8fda7fee35046f2def935086 37483f6be8a2df7a4392
x_2	179	wdlwk 45g	ab43bj3 6k34	SH A-2	586900065999e00dfd03caec2bd5eb43dd939f082d b4718edecd72fabfdcdbec
x_3	180	23nkd so4jf	u96kv9 6hk99	SH A-2	d2f483672c0239f6d7dd3c9ecee6deacbcd59185855 625902a8b1c1a3bd67440
x_4	181	mkfk5 68dk	r345l34 5cc4l	SH A-2	5d389f5e2e34c6b0bad96581c22cee0be36dcf627cd 73af4d4cccacd9ef40cc3
x_5	180	anv40 7aor	krbkuu6 3233	SH A-2	13671077b66a29874a2578b5240319092ef2a10432 28e433e9b006b5e53e7513
x_6	178	23nfro 86m	23vjd3o dodo	SH A-2	36ebe205bcdfc499a25e6923f4450fa8d48196ceb4f a0ce077d9d8ec4a36926d

Table 9 Operation of salted hash function with SHA-2

Peppered hashing provides additional protection to individuals in the way that it adds another layer of security for the prevention of re-identification. For instance, in the case of a data breach, an adversary who has gained knowledge of the pseudonyms must still need to brute-force the database if no disclosure of peppers has occurred. Therefore, peppered hashing seems to overcome some of the pitfalls of salted hashing.

4.3.1.1.2 Encryption

4.3.1.1.2.1 Symmetric Encryption

Symmetric encryption consists of the use of one secret key to both encrypt and decrypt electronic information. Parties relying on symmetric encryption must share the secret key to enable the decryption process. Symmetric encryption transforms the initial identifier (but also the complete dataset) into a pseudonym (or ciphertext), which is then decrypted to reveal the initial identifier. For this purpose, an encryption algorithm is used, such as AES.[151] In the following Table 10 and Table 11, the encryption and decryption processes using the AES encryption algorithm are exemplified.

Individual	Initial identifier *m* Height (cm)	Secret key *k* key size 32	Encryption algorithm *e*	Encrypted Output *e(k,m)* Output size 256
x_1	180	f1nd1ngn3m1ab43bj36k34u96kv96hk9	AES	E6kUjGZ1UF7gfcZpHrrTCg==

Table 10: Operation of encryption with AES

As can be seen above, the initial identifier *m* corresponding to the height of the individual x_1, together with the secret key *k,* is encrypted by one party using the encryption algorithm *e.* The resulting encrypted output is the pseudonym E6kUjGZ1UF7gfcZpHrrTCg==. This pseudonym is then decrypted by the other party by conducting the inverse process with the shared secret key *k.* The resulting decrypted output is 180, which corresponds to the height of individual x_1, as shown below.

Encrypted Output *e(k,m)* Output size 256	Encryption algorithm *e*	Secret key *k* key size 32	Decrypted output = Initial identifier *m* Height (cm)	Individual
E6kUjGZ1UF7gfcZpHrrTCg==	AES	f1nd1ngn3m1ab43bj36k34u96kv96hk9	180	x_1

Table 11: Operation of decryption with AES

The pseudonyms resulting from symmetric encryption satisfy the same properties as keyed hashing as long as different secret keys are used, no third party has access to the secret key, and state-of-the-art algorithms and sufficient lengths are used.[152] These are, (i) the reproducibility of the pseudonym by third parties is prevented as the generated pseudonyms are different depending on the shared secret key, therefore ensuring the unlinkability across domains; and (ii) any third party without the knowledge of the key would not be in a position to reveal the original identifier from the pseudonym. Aside from this, it is important to note that if the secret key is destroyed, it may not be possible to attribute the pseudonym to the initial identifier, even for the controller holding the initial identifier.[153] It remains, therefore, subject to debate whether the deletion of the secret key in symmetric encryption should be considered an anonymisation technique.

4.3.1.1.2.2 Asymmetric encryption

Asymmetric encryption consists of the use of two keys, a public and a private key, to both encrypt and decrypt electronic information. Parties relying on asymmetric encryption must rely on the public key to encrypt the data and on the private key to decrypt it. Whereas the public key can be used by anyone other than transacting parties, the private key must remain secret by each transacting party. Asymmetric encryption transforms the initial identifier (but also the complete dataset) into a pseudonym (or ciphertext), which is then decrypted to reveal the initial identifier, such as symmetric encryption does. In asymmetric encryption, public and private keys are mathematically related but appropriately

[151] The Advanced Encryption Standard (AES) is a specification standard established by the U.S. National Institute of Standards and Technology (NIST) in 2001. See FIPS, Federal Information Processing Standards Publication 197, "Advanced Encryption Standard ", 2001.
[152] ENISA, Algorithms, key size and parameters report – 2014, November, 2014, 36.
[153] ENISA, Recommendations on shaping technology according to GDPR provisions An overview on data pseudonymization, November 2018, 26.

distinguished by the introduction of randomness in the encryption process to prevent the determination of the private key. As a result, pseudonyms generated with the same public key are each time different without prejudice to their private decryption. Asymmetric encryption is commonly used in scenarios where the party in charge of pseudonymizing the data is not authorized to perform re-identification, e.g. in health contexts, or where tracking the data subject is not needed, and different pseudonyms can be attached to the same person. Consider Table 12 below, where the initial identifier m corresponding to the height of the individual x_1, together with the public key k_p, is encrypted by one party using the encryption algorithm e. The resulting output is the pseudonym $e(k_p,m)$.

Individual	Initial identifier m Height (cm)	Public key k_p key size 2048 bit	Encryption algorithm e	Encrypted Output $e(k_p,m)$ Output size 64
x_1	180	MIIBIjANBgkqhkiG9w0BAQEFAAOCAQ8AMIIBCgKCAQEAon6EwqgPjwySsc1GJ9Up7hp6EvtJO9F5/Hg7c+7v0aLILa83TldL+z9Vquw5RBcxeHBEXxiBC4twE4TOTV+8LBM86ukO2NMxfbi9bPfd3gSTtc8FVWzpm69Ytzoguzlyh6ir/g56DcZpaSgX43f+X6OWU1ZrnMUb1JAN8Q5nQoc6pWY2/ksyghrF0XImX1BgmKY9lSiDa5tB48B+Wdw53INhBHA94ydZKqaYUTaL9pHBdmh4yGqguwm7uEuPsZmTt6H47nTBlS73Y3BLmLwtE0wF5rD/VtD++J/+nQ4+X3EVaYVtXH6NyxmxxD+4TmcH3b2FdOSeyEuLSa+AH9l34vQIDAQAB	RSA	EaexuOzHRqc0c56ew97J/rkdkjaaY4B3WS+isq3Mb8s0ljF2VQuvfCTbzdG2HlOZkjaaS72vdRw9HnJ3kSqT6XED9BUUSHWlztRdQL1+YAyddiAHcZPs4NuaVPU8b+kSPk+66XPq6IlvKT2HDpKfdGa0bTczaHqyztA68CbhyMVkh6sl1ujGlKAwgnGGt4KkC3u5/yxszERqz4+slrZjPWOFa/cyOlUBIsVawiiH8YyyYrP2ZJqSKfmoHxtnUjZ+Y7EFY4S6pUMzWAr1PD2rc2zgVGCoa4L1t3UMqPq13DEOfQTNyQkUcrevUXlnGLo3am7j3CznXyQduD4Erb2N+g==	122

Table 12: Operation of encryption with RSA

The pseudonym $e(k_p,m)$ is then decrypted by the other party by using the private key k_t, which results in the decrypted output m=180 cm corresponding to the height of individual x_1, as shown in Table 13 below.

Encrypted Output $e(k_p,m)$ Output size 64	Encryption algorithm e	Private key k_t key size 2048 bit	Decrypted output = Initial identifier m Height (cm)	Individual
EaexuOzHRqc0c56ew97J/rkdkjaaY4B3WS+isq3Mb8s0ljF2VQuvfCTbzd	RSA	MIIEvgIBADANBgkqhkiG9w0BAQEFAASCBKgwggSkAgEAAoIBAQCifoTCqA+PDJKxzUYn1SnuGnoS+0k70Xn8eDtz7u/RosgtrzdOV0v7P1Wq7DlEFzF4cERfGIELi3AThM5NX7wsEzzq6Q7Y0zF9uL1s993eBJO1zwVVbOmbr1i3OiC7OXKHqKv+DnoNxmlpKBfjd/5fo5ZTVmucxRvUkA3xDmdChzqlZjb+SzK	180	x_1

G2HlOZkj aaS72vdR w9HnJ3kS qT6XED9 BUUSHWl ztRdQL1+ YAyddiA HcZPs4Nu aVPU8b+k SPk+66XP q6IlvKT2 HDpKfdG a0bTczaHq yztA68Cbh yMVkh6sl 1ujGlKAw gnGGt4Kk C3u5/yxsz ERqz4+slr ZjPWOFa/ cyOlUBIs VawiiH8Y yyYrP2ZJq SKfmoHxt nUjZ+Y7E FY4S6pU MzWAr1P D2rc2zgV GCoa4L1t 3UMqPq1 3DEOfQT NyQkUcre vUXlnGLo 3am7j3Czn XyQduD4 Erb2N+g= =		CGsXRciZfUGCYpj2VKINrm0HjwH5Z3Dn cg2EEcD3jJ1kqpphRNov2kcF2aHjIaqC7Cbu 4S4+xmZO3ofjudMGVLvdjcEuYvC0TTAX msP9W0P74n/6dDj5fcRVphW1cfo3LGbEP7 hOZwfdvYV05J7IS4tJr4Af2Xfi9AgMBAAE CggEAdsyc71vDlEgS798uDW0xLJ065qnFD vDFiKm2fboQp2mlhJD3lFUKffP+A+qUq41 xY6Zgtk3J+tDE7eBInBUEFA00mmItsqQlK M9p3kMReIJMI73pHl10JZh6+eqLh5Ymf7v 3ktSus2d+JZ1kaa+O0AlfnCownvsUr8FqD/U B1Yaus3iawoxjWs1laBxbdEvRIA7nAlQ0GF mfNJj13EsLK/Hw6P71dC7VfPgEl6Y3DGdl B9ukcAGm4CP9FJiSP9K4lTqdUoxFJ/TZPlu imRcYP1QzjGWSaCwNguWmbv5M2gdPH Z5ij79WOvy748qYGyhRPu9vCcgSOeIdFZp utf7kQQKBgQDuCl5VWeoJvlTqnStspa/8xy bxVUuR/R2garVvghfqml6fJXTWDbhrZa4S T+5bxWS43qq8ZHxpQHSkyLEUOlvIAPUg 7oPxqYMahPAsWkpYJ1uXqEPRCcxFn/CsR aUgQuTUAwpeKCGa4EL9wt1LWV5eB+0r xNIdnbFQ6mbHFA6CVQKBgQCuwQMKH SI936B0bkI9YaoxICJN4LiGCUa3/1H6LjMU 4x6QzrVQoPQpR55AhxMhJKZXiGdGgaA8 1WLDegTFAOFRjbs+yoOHrC/Y9l4txdMpY UnSzgmuVDUh5urgk6y3hqV9PyClUXCRc0 fZB6Miusg29bBoam7tNuHx4RfGmp8UyQK BgQC0YycZhwnUWGgBYxmFPAohhMn+G KUr/KR27GaSGgQFxPXvpHLlw3/94xAG0I vQ/8VUU2kPWxbsq+u49F089vyWlCMu0Z ErkUnIzvUypzmvlCPUya6autIR8SJxalW4H PlLsQIqWD8fHtOlegY5E4BvwZH5mS2hQ9 8bcoZwVsKoAQKBgQCsMmCOXLrbATqY v/ThixUrJyDmYGMzIHzg3eOnhFKtaEc8JZ OgiZlN+9ZCe1csN3L5md06Koz8pL+XIusE PKPJusEhVGhbDh5vygRvUhmLEuStpnz/nN ZmO6aB+MIebb0wNz4x6JflmxTXFKF5nVe gYGSd3xLDCGuH7meBOec7kQKBgDMnN m74IF07SSxAt6skDEzUp8hiZ0UmeVDNCZ 0NZT2qKnAmmygmpPDk1QNaJyaC5MuC7 hYkwhkizUoMBnYgt5oFi0vKYqTPy1T3fT wq9eU3oaQRUHOVn2jPmFFjnlaLdGt6VvP BI5skXZNLMu5Ylq6ZN805PfJLjEiSkdV7W Kjz		

Table 13: Operation of decryption with RSA

It should be stressed that asymmetric encryption necessitates the usage of very large keys. For instance, NIST has recommended a minimum of 2048-bit keys for RSA since 2015.[154] As a result, the computation of asymmetric encryption is less efficient than symmetric encryption. It has also been argued that most asymmetric encryption algorithms may not offer strong security in the post-quantum era.[155]

[154] Barker, E., & Dang, Q. (2014). Draft NIST special publication 800-57 part 3 revision 1. Recommendation for Key Management.
[155] Mavroeidis, V., Vishi, K., Zych, M. D., & Jøsang, A. (2018). The impact of quantum computing on present cryptography. arXiv preprint arXiv:1804.00200.

4.3.1.1.3 Other cryptography-based techniques

4.3.1.1.3.1 Homomorphic encryption

Homomorphic encryption allows computation on encrypted data. Computing on encrypted data refers to the fact that a party P_n having the initial identifiers or input m_n and wanting to calculate the function f to obtain $f(m_1,\ldots,m_n)$, can instead compute the encryptions or pseudonyms of the inputs c_n to obtain $f'(c_1,\ldots,c_n)$, which can be decrypted to $f(m_1,\ldots,m_n)$.The benefit of homomorphic encryption is that personal data remains confidential while being analyzed or mined without the need to decrypt it and compromise the output. This approach to computation is very useful as it allows, for instance, third parties cloud computing providers to perform analysis on data without disclosing the initial values and send the results in an encrypted format, which can then be decrypted by the issuing party.

The computation of the ciphertext resulting from the encryption is possible due to the fact that usually, the initial plaintext embodies a certain algebraic structure which can be replicated in the ciphertext. For instance, standard homomorphic encryption schemes typically restrict the function f to be an algebraic operation associated with the structure of the plaintexts. Therefore, if the plaintext space corresponds to a certain group G, the associated ciphertext may be the product $G \times G$ while f is restricted to the group operation on G.[156] In the case of fully homomorphic encryption (FHE), the advance relies upon the extension of the function f to be any function so that a party can apply any arithmetic circuit to the encrypted data and obtain an output ciphertext that encrypts the output that would be obtained if the circuit was directly applied to cleartexts.[157] Of course, to render a scheme fully homomorphic with respect to a functionally complete set of operations and iterate those operations from that set comes with several drawbacks, including scalability among multiple parties, computational overhead, and secret function evaluation.[158]

Consider the following
explaining the process of homomorphic encryption.[159]

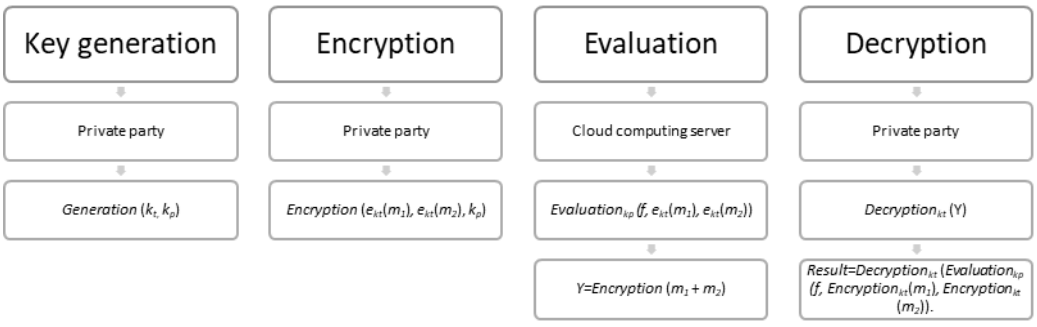

Figure 5: functions of homomorphic encryption, extracted from Parmar et al., 2014

As shown above, homomorphic encryption encompasses four different functions, where:

- first, in the key generation function, the party will generate a public key k_p and a private key k_t to encrypt the initial identifiers m_1 and m_2, *i.e. Generation* (k_t, k_p);

[156] Armknecht, F., Boyd, C., Carr, C., Gjøsteen, K., Jäschke, A., Reuter, C.A., & Strand, M. (2015). A Guide to Fully Homomorphic Encryption. IACR Cryptol. ePrint Arch., 2015, 1192, p. 2. See also Taher El Gamal.(1985) A public key cryptosystem and a signature scheme based on discrete logarithms. IEEE Transactions on Information Theory, 31(4):469–472.
[157] ENISA, Privacy and Data Protection by Design – from policy to engineering, December 2014, 43.
[158] Armknecht, F., Boyd, C., Carr, C., Gjøsteen, K., Jäschke, A., Reuter, C.A., & Strand, M. (2015). A Guide to Fully Homomorphic Encryption. IACR Cryptol. ePrint Arch., 2015, 1192, 9-10.
[159] Cf. Parmar, P., Padhar, S.B., Patel, S.N., Bhatt, N.I., & Jhaveri, R.H. (2014). Survey of Various Homomorphic Encryption algorithms and Schemes. *International Journal of Computer Applications, 91*, 26-32.

- second, in the encryption function, the party will encrypt, with the private key k_t, the initial identifiers m_1 and m_2 to generate a ciphertext which, along with the public key k_p, will be sent to the server for evaluation, *i.e. Encryption* ($e_{kt}(m_1)$, $e_{kt}(m_2)$, k_p);
- third, in the evaluation function, the server will use the public key k_p to compute the function f on the ciphertext for evaluation purposes, *i.e.* $Evaluation_{kp}$ (f, $e_{kt}(m_1)$, $e_{kt}(m_2)$), and send it to the party in the encrypted format $Y=Encryption$ ($m_1 + m_2$); and
- fourth, in the decryption function, the evaluation function $Evaluation_{kp}$ (f, $e_{kt}(m_1)$, $e_{kt}(m_2)$) will be decrypted by the party using the private key k_t to arrive to the computed result $Decryption_{kt}$ (Y), this being: $Result=Decryption_{kt}$ ($Evaluation_{kp}$ (f, $Encryption_{kt}(m_1)$, $Encryption_{kt}$ (m_2)).

It has been postulated that homomorphic encryption allowing full arithmetic operations on encrypted data, is, at best, secure against known-cleartext attacks.[160] These attacks consist of having access to both the plaintext and the ciphertext to reveal further secret information, such as secret keys and code books. Therefore, homomorphic encryption is still seen as a pseudonymization technique.[161]

4.3.1.1.3.2 Secure multiparty computation (MPC)

Multiparty computation is a technique that, similar to FHE, deals with protocols which allow a set of parties to jointly compute a function of their inputs or identifiers while avoiding revealing anything but the output of the said function.[162] MPC allows the input of the parties to remain generally secret during the whole processing of data aggregation, thus being considered a sophisticated privacy-preserving tool for pseudonymization.[163] MPC protocols can vary in their efficiency, security or robustness, and they can be set up for different scenarios[164] and technological schemes, such as homomorphic encryption,[165] garbled circuits,[166] oblivious transfers[167] and secret sharing.[168] In some cases, such as homomorphic encryption, MPC has been considered by several authors to provide anonymization guarantees.[169]

MPC works as follows: the initial identifiers are encrypted by randomly splitting them into secret shares. These secret shares can be distributed among P_n parties, so that any subset of $t + 1$ or more of the parties can reconstruct the secret, yet no subset of t or fewer parties can learn anything about the secret. The secret shares can thus be distributed among the P_n parties, which jointly perform the necessary computations $f(x_1,x_2,x_3) = (y_1,y_2,\ldots,y_n)$ without revealing the initial identifiers x_n.[170] The number of parties may vary depending on the required trusted model, but privacy and confidentiality are ensured as long as a subset of the trustee acts honestly or, in other technical words, the trusted parties do not collude.

[160] ENISA, Privacy and Data Protection by Design – from policy to engineering, December 2014, 43.
[161] ENISA, Recommendations on shaping technology according to GDPR provisions An overview on data pseudonymization, November 2018, 27.
[162] Spindler, G., & Schmechel, P. (2016). Personal data and encryption in the European general data protection regulation. J. Intell. Prop. Info. Tech. & Elec. Com. L., 7, 163.
[163] European Union Agency for Cybersecurity, Data Pseudonymisation: Advanced Techniques & Use Cases. Technical analysis of cybersecurity measures in data protection and privacy, 23. See also European Union Agency for Cybersecurity, Recommendations on shaping technology according to GDPR provisions An overview on data pseudonymisation, 1-42.
[164] The existing work in MPC protocols include, inter alia, decision tree models, linear regression models, and neural network architectures.
[165] Homomorphic encryption is a form of encryption that allows the performance of computations on encrypted data without decrypting it. See Armknecht, F., Boyd, C., Carr, C., Gjøsteen, K., Jäschke, A., Reuter, C. A., & Strand, M. (2015). A guide to fully homomorphic encryption. IACR Cryptology ePrint Archive, 2015, 1192.
[166] Garbled circuits is a form of encryption that breaks down a function into a Boolean circuit to allow a finer grained manipulation of the function. See Yakoubov, S. (2017). A Gentle Introduction to Yao ' s Garbled Circuits. Boston University, 3.
[167] Oblivious transfer is a secure computation protocol in which the sender transfers one of potentially many pieces of information to a receiver, but remains oblivious as to what piece (if any) has been transferred. See Kilian, J.: Founding cryptography on oblivious transfer. In: Proceedings of the 20th Annual ACM Symposium on Theory of Computing, 2–4 May 1988, Chicago, Illinois, USA, pp. 20–31 (1988).
[168] A secret-sharing scheme is a cryptographic method by which a dealer distributes shares to parties such that only authorized subsets of parties can reconstruct the secret. See Beimel A. (2011) Secret-Sharing Schemes: A Survey. In: Chee Y.M. et al. (eds) Coding and Cryptology. IWCC 2011. Lecture Notes in Computer Science, vol 6639.
[169] Scheibner, J., Raisaro, J. L., Troncoso-Pastoriza, J. R., Ienca, M., Fellay, J., Vayena, E., & Hubaux, J. P. (2021). Revolutionizing medical data sharing using advanced privacy-enhancing technologies: technical, legal, and ethical synthesis. Journal of medical Internet research, 23(2), e25120. See also: Veeningen, M., Chatterjea, S., Horváth, A. Z., Spindler, G., Boersma, E., van der SPEK, P., ... & Veugen, T. (2018). Enabling Analytics on Sensitive Medical Data with Secure Multi-Party Computation. In MIE (pp. 76-80).
[170] Zhao, C., Zhao, S., Zhao, M., Chen, Z., Gao, C., Li, H., & Tan, Y. (2019). Secure Multi-Party Computation: Theory, practice and applications. *Inf. Sci., 476*, 357-372.

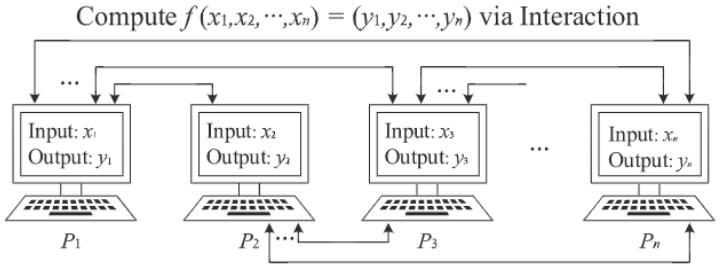

Figure 6: operation of MPC, based on C. Zhao et al., Information Sciences 2019

While the theoretical assumptions relating to the security guarantees of MPC can vary depending on the type of actors involved in the protocol, the desired security levels, and the types of output guarantees,[171] these can be taken into account to steer the debate on whether MPC should be considered as a robust pseudonymization or anonymization technique for the computation and sharing of personal data.
According to the extensive research performed by Escudero,[172] three main scenarios can be extracted as fundamentally determinants for the legal categorization of MPC as an appropriate pseudonymization or anonymization technique. This implies that, from a data protection perspective, the qualification of MPC would be subject to a probabilistic parameter measuring the risk of re-identification of its random fragments of personal data.

- In the first setting involving both active and passive adversaries where said adversaries are a minority $t < n/3$, privacy and output delivery may be perfect. In such cases, it could be argued that MPC would provide perfect security as postulated by information security theory, therefore sufficiently preventing unlawful access to personal data. Under this assumption, it could be questioned whether the resulting pseudonyms could be considered anonymous.
- In the second setting, where $t < n/2$, statistical security may be possible with full output guarantees. In these cases, a case-by-case analysis that considers, inter alia, the statistical security parameter, together with the access to the executions of the protocol, could offer an answer as to the quality of MPC fragmented data as pseudonyms or non-personal data.
- In the third setting involving $t < n$, computational complexity would be the parameter defining the nature of the processed personal data. In these cases, MPC may be considered a pseudonymisation technique. It is, therefore, foreseeable that additional safeguards should be in place where computational security offers the highest security settings.

A point of concern in this area is whether the division of the data into fragmented strings should be treated as a traditional form of encryption and if the resulting data 'chunks' should be treated as personal data within the meaning of the data protection framework. Here, it seems that the categorization of both MPC and fragmented data is still a debatable issue.[173] On the one side, it has been argued that MPC differs from traditional encryption in the way that more than a single key is needed to decrypt the data, as secret shares are distributed among different entities with strong interests in keeping the data confidential. On the other side, it has also been put forward that data fragments themselves do not contain information regarding a natural person and *should not be seen as personal data*; only if all fragments of the data were gathered and put together would be considered as such. This interesting notion calls for a wider debate on the fragmentation of personal data as a way of anonymization, particularly in relation to the element of 'relating to' in the definition of personal data. Such a debate would, in any case, be contingent upon the adoption of a relative approach to data protection. Contrarily,

[171] Independence of the inputs by the parties and correctness of the outputs are also properties which a MPC protocol must ensure. See Yehuda Lindell, Secure multiparty computation, fn. 38, 1.
[172] Escudero, D. Multiparty Computation over Z/2 kZ.. Aarhus University. 2020. <https://www.escudero.me/pdfs/phd_thesis.pdf>.
[173] Damiani, E. D31.3. Evaluation and integration and final report on legal aspects of data protection. Privacy-Preserving Computation in the Cloud (PRACTICE). November 2016. ICT-609611, available at: https://practice-project.technikon.com/downloads/publications/year3/D31.3-Evaluation-and-integration-and-final-report-on-PU-M36.pdf. See also <https://medium.com/applied-mpc/simplified-gdpr-compliance-using-mpc-cryptography-un-and-ec-studies-explain-b2c21ecd0d7b>.

if an absolute approach were to be argued, one could still oppose the consideration of personal data to the data fragments resulting from the MPC.

4.3.1.1.3.3 Tokenization

Tokenization this advanced pseudonymization technique consists of replacing identifiers with randomly-generated values, known as tokens, without any mathematical relationship and without altering the type or length of the data.[174] This is an important difference with respect to encryption. As opposed to the latter, the invariability of data types and lengths in tokenization prevents any unintelligibility of information through its processing in intermediate systems. At the same time, it also implies a decrease in the computational resources needed to process the tokens. Since there is no involvement of keys or algorithms to derive the original identifier from the token, the knowledge of a token does not imply the disclosure of personal data. The relationship between the pseudonym or token and the initial identifier is usually stored in a database, or token vault, which is secured, often via encryption. The most prevalent uses of tokenisation can be found in the financial sector, where tokens are used to protect payment card data. This can be better explained in Table 14 below.

Individual	Initial identifier *m* Height (cm)	Token *t*	Token vault	
x_1	180	432	Initial identifier *m*	Token *t*
			180	432

Table 144: Operation of tokenization

In the tokenization process, the initial identifier *m* is replaced by the randomly-generated token *t.* Different methods can be used to generate the token, including random number generation,[175] encryption, or hashing. The token *t* can then be used in various applications. Should the initial identifier *m* be retrieved, for instance, for a credit card payment, the token *t* is submitted to the token vault, which contains the relationship between *m* and *t.* After validation, the transaction is authorized.

It is apparent that the random nature of tokens satisfies both the properties of unlinkability across domains and the revelation of the original identifier. It should be noted, however, that re-identification is still possible for the data controller storing the relationship between the token and the initial identifier. Despite its efficiency as a pseudonymization technique, one of the pitfalls of tokenization is the difficult synchronization of tokens across several systems, which may need, in many cases, the use of several applications.

4.3.1.2 Quantum Computing

Messages and data are encrypted by cryptographic algorithms that provide sufficient safeguards to allow network communication and transactions.[176] At present, two primary forms of encryption exist, namely symmetric encryption and asymmetric encryption.[177] In the former, the sender and the receiver use the same secret key and the same cryptographic algorithm to encrypt and decrypt data. Symmetric encryption is typically used where speed is the priority over increased security, *e.g.,* for transactions in the banking sector.[178] In the latter, the keys come in pairs. Each party has its own private and public key, and only the person who owns the private key can decrypt the message. Asymmetric encryption is

[174] ENISA, Recommendations on shaping technology according to GDPR provisions An overview on data pseudonymization, November 2018, p. 28.
[175] Random number generation (RNG) is a technique where random numbers are assigned to the initial identifiers. RNG provides strong data protection unless the mapping table is compromised. The pitfalls of RGN are the possibility of collisions and scalability. ENISA, Data Pseudonymisation: Advanced Techniques & Use Cases. Technical analysis of cybersecurity measures in data protection and privacy. January 2021, 13.
[176] <https://www2.deloitte.com/us/en/insights/topics/cyber-risk/crypto-agility-quantum-computing-security.html>.
[177] There is another form of encryption called hashing, where a cryptographic algorithm is used to transform large random size data to small fixed size data.
[178] Mavroeidis, V., Vishi, K., Zych, M. D., & Jøsang, A. (2018). The impact of quantum computing on present cryptography. arXiv preprint arXiv:1804.00200.

typically used where increased security is the priority over speed, e.g. for internet communication, digital certificates and signatures or blockchain applications.

Cryptographic algorithms play hence a fundamental role in the information technology infrastructure and communications, as they ensure confidentiality, integrity, authenticity and non-repudiation in data transmissions and storage.[179] As a disruptive technology using quantum mechanical phenomena to concurrently perform large sets of processing tasks, quantum computing may put at risk the encryption algorithms under which information technology infrastructure and communications rest.

Quantum computing possesses certain characteristics derived from quantum mechanics that make it possible to solve complex factorization problems with which traditional computers struggle. Instead of working with bits, quantum computers work with quantum bits or *qubits*. Qubits are able to simultaneously take a value of 0 or 1, as opposed to traditional bits, which have a single state of 0 or 1. This enables quantum computers to perform multiple parallel calculations for which conventional computers are not suited.[180] As a result, certain authors have claimed an alleged 'supremacy' of quantum computing over conventional computing, which would allow for the cracking of the present cryptography.[181]

Although the impact of quantum computing on the described cryptographic techniques may vary according to their nature and properties,[182] it has been claimed that quantum computers may put at risk the entire system network.[183] This becomes more evident in the current internet communication infrastructure. Since many popular internet, asymmetric encryption protocols rely on mathematical calculations which break down large numbers into their prime factors, such as HTTPS, SSL or TLS,[184] practical quantum computing may expose their vulnerabilities and perform undetected modifications.

The principle of privacy by design aims at consistently safeguarding fundamental rights and freedoms *vis-à-vis* the dynamism of technological development by integrating privacy settings throughout the whole engineering process. As such, the principle of privacy by design may be affected by quantum computing in the way that privacy settings established by controllers and processors by design may fail to provide sufficient safeguards if they cannot keep up with the development of quantum computing capabilities.[185] The principles of integrity and confidentiality mandate safeguarding personal data through security measures, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical and organisational measures and procedures. For these reasons, technical and organizational measures to ensure the integrity and confidentiality of network systems are of the utmost importance to minimize the risks of quantum technologies. If data controllers or processors use encrypted protocols which do not secure the *minimum* standards of protection for the processing of personal data or are openly vulnerable to being hacked, potential breaches of the principles of integrity and confidentiality may occur, thus putting at risk the

[179] Vigil, M., Buchmann, J., Cabarcas, D., Weinert, C., & Wiesmaier, A. (2015). Integrity, authenticity, non-repudiation, and proof of existence for long-term archiving: a survey. Computers & Security, 50, 16-32.
[180] Bruno, L., & Spano, I. (2021). Post-Quantum encryption and privacy regulation: can the law keep pace with technology?. Eur. J. Privacy L. & Tech., 72. 2021, p. 2.
[181] Arute, F., Arya, K., Babbush, R. et al. Quantum supremacy using a programmable superconducting processor. Nature 574, 505–510 (2019). https://doi.org/10.1038/s41586-019-1666-5
[182] Quantum computing theory implies a threat to present modern asymmetric cryptography whose security is based on the difficulty of factorizing large prime numbers. Likewise, symmetric cryptography can also be affected by specific quantum algorithms. On a general note, symmetric encryption offers however more resistance to quantum computers.
[183] Mavroeidis, V., Vishi, K., Zych, M. D., & Jøsang, A. (2018). The impact of quantum computing on present cryptography. arXiv preprint arXiv:1804.00200.
[184] Transport Layer Security (TLS) is the successor of the Secure Sockets Layer (SSL) as a cryptographic protocol designed to provide communications security over a computer network. The protocol is widely used in applications such as email, instant messaging, and voice over IP, but its use in securing Hyper Text Transfer Protocol Secure (HTTPS) remains the most publicly visible. HTTPS appears in the URL when a website is secured by a TLS or SSL certificate.
[185] Chen, L., Chen, L., Jordan, S., Liu, Y. K., Moody, D., Peralta, R., ... & Smith-Tone, D. (2016). Report on post-quantum cryptography (Vol. 12). Gaithersburg, MD, USA: US Department of Commerce, National Institute of Standards and Technology.

fundamental rights and freedoms of individuals in relation to the processing of their personal data and beyond.

Although the validity and consequences of 'quantum supremacy' are still debatable under the current start of the art,[186] certain authors have predicted that, if readily available to cyber adversaries, quantum computing will be able to break the security of 'nearly all modern public-key cryptographic systems.'[187] European lawmakers and data protection authorities have shown certain concern about quantum computing perils and have already expressed their apprehensions as to the protection of personal data.[188] Thereby, a new generation of cryptographic systems that are secure against both quantum and classical computers and can interoperate with existing communications protocols and networks is being brought about. This new kind of cryptography has been called 'post-quantum cryptography' or 'quantum-resistant cryptography.' Current efforts in this area include the standardization process made by the National Institute of Standards and Technology (NIST), aiming at delivering asymmetric quantum-resistant algorithms.[189]

The adoption of quantum-resistant cryptography may, notwithstanding, be a slow and burdensome process. It would require not only stored keys and data be re-encrypted with said quantum-resistant algorithms but also backups to be either deleted or physically secured. In addition to this, the introduction of quantum-resistant algorithms may also require the implementation of appropriate technical and organisational measures to avoid innocuous non-identifying information becoming subversive identifiable information. Appropriate technical measures may include the replacement of cryptographic libraries, implementation validation tools, operating systems, hardware, devices and protocols. Appropriate organizational measures may consist of user and administrative procedures, security standards, and best practice documentation. Of course, such an undertaking would equally require high expenditures in terms of time, cost, and resources. Significant public engagement to assure trust in the selected algorithms would also be required.

4.3.2 Results from interviews

This section will discuss some of the main findings gained through the interviews conducted for this study. The full interview reports may be found in the annex to this report.

From the interviews, the following insights were drawn with respect to pseudonymisation and encryption:

- Pseudonymisation and pseudonyms: Technical experts agree that pseudonymization means replacing one or more identifiers with a pseudonym. This is in line with the legal definition of the GDPR. However, the definition in the GDPR does not refer to the word pseudonym. Pseudonymization is defined there as processing where additional information that allows re-identification is stored in a different place; there can be pseudonymous data without having an explicit pseudonym. This can be, e.g., illustrated by the famous example of the hospital in the USA years ago where data such as age, diagnosis, address data etc. of patients were available online because they were assumed to be anonymous. However, this turned out to be not the case; as with other public information, the researcher managed to re-identify many of them. In this example, there was no use of explicit pseudonyms. But at the same time, these data could be considered pseudonymous data under the GDPR, as they are certainly not anonymous data. In that light, the two definitions are very close but not fully identical.

[186] Mattsson, J. P., Smeets, B., & Thormarker, E. (2021). Quantum-Resistant Cryptography. arXiv preprint arXiv:2112.00399.
[187] <https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04282021.pdf>.
[188] The European Data Protection Supervisor (EDPS), 2020, TechDispatch #2/2020: Quantum Computing and Cryptography, available at: <https://edps.europa.eu/data-protection/our-work/publications/techdispatch/techdispatch-22020-quantum-computing-and enEur≥.
[189] Chen, L., Chen, L., Jordan, S., Liu, Y. K., Moody, D., Peralta, R., ... & Smith-Tone, D. (2016). Report on post-quantum cryptography (Vol. 12). Gaithersburg, MD, USA: US Department of Commerce, National Institute of Standards and Technology.

- Pseudonymisation and anonymisation: While the legal domain distinguishes between pseudonymous and anonymous data, from a technical perspective, it is more of a scale.
- Pseudonymisation relates to risk management: Technical experts generally aim at preventing harm but do not focus on legal categorisations when doing so, but instead focus is on the contextual actors, likely outcomes and threats. Pseudonymisation is generally used as a way to reduce risk while retaining personal data, while anonymisation is used when the data are no longer necessary within an organisation.
- Pseudonymisation and encryption: Pseudonymisation is not the same as encryption, for pseudonymisation, the recipient of the data cannot go back to the main text or the original data; only with the use of additional information might this be the case.
- Pseudonymisation and aggregation: pseudonymisation is popular to use on statistical or aggregate data. When an actor conducting statistical analysis needs to have access to the original data in order to perform the analysis, pseudonymisation can be an important asset to allow such processing and, at the same time, ensure a high level of protection.
- Standardisation and certification: no matter what technique or technology is used, the standardization of techniques and examining new proposed techniques are deemed to be important. In addition, certification of new techniques could be conceived. For example, a new form of encryption may not offer the best form of encryption. Thus, it may be safer to focus on encryption algorithms that have already proven themselves and are standardized, such as AES or RSA algorithms. If, however, the new encryption techniques were assessed and certified, such would make it easier for data controllers.
- Hashing: for hashing, one cannot exclude a scenario in which it can be useful, as there could be a scenario with very low data protection risks. However, the use of classic hashing (i.e. the controller does not have a secret key) for pseudonymization has significant drawbacks, as someone having knowledge of the pool of input may quite easily re-identify some people from the pseudonymized dataset.
- Transit: some technical experts see data as having two states: Data can either be at rest or data can be in transit. Data protection and cyber security generally focus on data in transit because it is a more imaginable risk that data get intercepted. However, data is at rest for most of its lifetime. Storage encryption is perhaps the most important form of encryption.
- Quantum computing: Encryption is always a race against the clock in cybersecurity where technological possibilities increase, so cybersecurity has to improve as well. One of the challenges in terms of technology is quantum computing, as data still need to be adequately protected in 20 years' time.
- Privacy enhancing cryptography: zero knowledge proofs, secure multiparty computation/secret sharing, homomorphic encryption etc. can be important tools for pseudonymization.
- Multiparty Computation (MCP): MPC could be used to some extent for degrees of anonymization and legal terms, most likely for pseudonymisation. The most commonly used in MPC is secret sharing. MPC offers different models or levels of strictness. For example, active security is the strictest form. MPC does not offer privacy guarantees for the output of the processing; the output can contain personal data. MPC protects input and intermediate results. Using MPC often includes respect for several GDPR standards, such as forms of data minimization, decentral processing, adhering to the purpose limitation, etc.

4.3.3 Results from workshop

The workshop held for this study yielded the following results:

- The definition/concept: The definition of pseudonymous data as it currently stands under the GDPR can be challenged and is unclear to many experts.

- Purpose of pseudonymization: pseudonymization aims to reduce risk. The question is how to assess the risks. There may be several risks involved with data processing and several ways to measure and assess those risks as well as several ways to mitigate those risks.
- Special legal status of pseudonymisation: It is unclear to many technical experts why pseudonymous data is explicitly mentioned in the GDPR, while other technical safeguards are not. It is important to consider whether pseudonymization is the best boundary marker for a lighter legal regime and if pseudonymisation should be the only means that is explicitly recognised under the GDPR.
- Pseudonymisation techniques: Hashing is often conceived as a way to achieve pseudonymisation, however, it is a very weak form of pseudonymisation. Another technique that could be more prominent in the future is the use of synthetic data. For pseudonymisation techniques, it is important to assess whether they actually offer protection to prevent abuse of legal exceptions. At least for every technique, the baseline is that the protective technique should be conducted properly. If applied properly, it can offer protection, if not, then it can lead to abuse of the technology.

4.4 Analysis

Five main tensions between the legal and the technical realm have emerged from this chapter:

1. The legal regime attributes a special status to pseudonymised data. When data are pseudonymised, a number of obligations do not apply or are applied less strict. Hence, pseudonymous data are conceived as an intermediate category between non-personal data and personal data, though it is clear that the GDPR applies to pseudonymous data. From a technical perspective, on the one hand, this choice is lauded.[190] The black-and-white approach taken in the data protection framework between personal data (fully protected) and non-personal data (processing restrictions are dissuaded) does not align with the more fluid and contextual understanding of anonymity by technical experts. Thus, the idea of an intermediate category is well received. On the other hand, it is unclear from a technical perspective why pseudonymisation, as a technique, should have a preferred and privileged position in the data protection framework and why other privacy-enhancing techniques are not put on the same level. In addition, the precise delineation between anonymous and pseudonymous, and pseudonymous and non-pseudonymous personal data is not always clear from a technical perspective.
2. Pseudonymisation is a state or outcome in which the data cannot be attributed to an individual without the use of additional information; the process for doing so can be achieved with various techniques.[191] The EU legislator equates encryption and pseudonymization on multiple occasions, or at least refers to pseudonymisation where it could have also referred to encryption, for example, in the context of Articles 25 and 40 GDPR. From a technical perspective, these are clearly distinct: pseudonymization aims at decreasing linkability, while encryption focuses on the confidentiality of information.
3. The GDPR is neutral in terms of which form of pseudonymisation or which pseudonymisation technique is used, but this choice may be challenged because some are clearly better than others.[192] For example, hashing is agreed to be weak.[193] Not only can pseudonymization easily be undone, but sometimes, it can also be used as a façade by actors who are either unwilling to delete identifiable information in their database or abuse it as an argument to avoid implementing further costly technical, technological, and organisational measures. That is why

[190] Esayas S. Y. (2015).The role of anonymisation and pseudonymisation under the EU data privacy rules: beyond the all or nothing‘ approach‖, in European Journal of Law and Technology, 6 (2).
[191] Mourby, M., et al. (2018). Are ‘pseudonymised’data always personal data? Implications of the GDPR for administrative data research in the UK. Computer Law & Security Review, 34(2), 222-233.
[192] Marx, M., Zimmer, E., Mueller, T., Blochberger, M., & Federrath, H. (2018). Hashing of personally identifiable information is not sufficient. SICHERHEIT 2018.
[193] ENISA, Recommendations on shaping technology according to GDPR provisions. An overview on data pseudonymization, November 2018.

there is discussion over which technologies should be used in different contexts; the legal regime provides no guidance on this point.

4. Often one pseudonymisation technique is not enough to prevent abuse.[194] It can be argued that pseudonymization is more than just a technical measure; to achieve the result as defined under the GDPR, it also requires organisational measures, such as the management of access rights for the personnel that has access to the key of the pseudonymised data.[195] The question is whether the technical protection is solid enough to protect against the harm of identification to warrant a lighter regulatory regime for pseudonymous personal data.
5. While, under the GDPR, pseudonymized data are considered personal data, some propose that is not the case for other jurisdictions and that under most legal regimes beyond Europe, including the USA, pseudonymised data are not considered to be personal data.[196] Nonetheless, there are non-EU jurisdictions, such as India, where pseudonymous data are seen as personal data. While under the GDPR it is clear that pseudonymous data are personal data, some scholars argue that the UK Information Commissioner's Office ('ICO') takes a different stance towards pseudonymous data. The ICO stresses that pseudonymisation can produce anonymised data on an individual-level basis, even though it may pose a greater privacy risk than aggregated anonymous data. The ICO uses the concept of pseudonymous data to be key-coded, referring merely to a de-identification technique for individual-level data. Therefore, it might be preferable to use the term 'de-identification' rather than pseudonymization to distinguish it from the specific GDPR definition.[197] However, on the other hand, in its newest guidance, the ICO very clearly states that it deems pseudonymous data to be personal data.[198] This illustrates the uncertainty and unclarity that still shrouds the idea of pseudonymous data.[199] When parties operate both in the EU and, for example, in the UK, they may have to take different technological and organisational measures and/or may be faced with different legal interpretations over the same pseudonymisation technique.

[194] ENISA, Data pseudonymisation: advanced techniques & use cases. Technical analysis of cybersecurity measures in data protection and privacy. January 2021.
[195] Jasmontaite, L., Kamara, I., Zanfir-Fortuna, G., & Leucci, S. (2018). Data protection by design and by default: framing guiding principles into legal obligations in the gdpr. European Data Protection Law Review (EDPL), 4(2), 168-189.
[196] Malekian, H. (2017). Pseudonymisation under the General Data Protection Regulation: A win-win approach?. In the Journal of Data Protection & Privacy, 1(3).
[197] Mourby, M., et al. (2018). Are 'pseudonymised'data always personal data? Implications of the GDPR for administrative data research in the UK. Computer Law & Security Review, 34(2), 222-233.
[198] ICO: Anonymisation, pseudonymisation and privacy enhancing technologies guidance, available at: <https://ico.org.uk/media/about-the-ico/consultations/4019579/chapter-3-anonymisation-guidance.pdf>.
[199] On the ICO's confusion see also: Finck, M., & Pallas, F. (2020). They who must not be identified—distinguishing personal from non-personal data under the GDPR. International Data Privacy Law.

Chapter 5: Sensitive and non-sensitive personal data

5.1 Introduction

Chapters 2, 3 and 4 form the core of this study. They relate directly to the research questions outlined in section 1.3. Chapter 5 discusses an additional distinctions between categories of data, namely the distinction between non-sensitive and sensitive personal data. This distinction is part of the data protection regime and is included because many of the problems, discussions and challenges as found in chapters 2, 3 and 4 also apply to these distinctions. Alternative approaches to defining the difference between non-personal and personal data, for example, necessarily also have a bearing on the distinction between non-sensitive and sensitive personal data. Hence, the one cannot be discussed without the other.

This chapter will assess the boundary between ordinary data and sensitive data and between sensitive personal data and non-personal data. For example, a question is whether a template for biometric data used for facial recognition is already sensitive personal data or still qualifies as non-personal data as the data cannot be used to identify a person. Another question is whether body tissue itself is to be understood as sensitive personal data or whether it should be considered non-personal data, as, without any additional identifying information, it does not enable a party to identify a person. It will discuss the categories that are considered 'sensitive' under the European data protection framework and assess different techniques that can be used for inferring sensitive data from non-sensitive data and anonymising sensitive data, in particular in the health care context.

Section 5.2 discusses the legal distinction between sensitive and non-sensitive personal data, section 5.3 describes the main techniques available for inferring sensitive data from non-sensitive personal data and non-personal data and section 5.4 analyses the gap between the legal regulation and technical reality.

5.2 Legal regulation

This section will give an overview of the legislative history of European data protection law as far as relevant for sensitive data. It will start with a discussion of Resolution 1973 (section 5.2.1), Resolution 1974 (section 5.2.2) and Convention 108 (section 5.2.3), all Council of Europe instruments. These are important because the concept of sensitive data was largely derived from the European Convention on Human Rights, especially articles 8 and 14, and the jurisprudence of the European Court of Human Rights. Subsequently, it will discuss how sensitive data are regulated in the EU, starting with the Data Protection Directive (section 5.2.4), the GDPR (section 5.2.5) and the Law Enforcement Directive (section 5.2.6), and then turning to the jurisprudence of the CJEU (section 5.2.7) and then turning to and an opinion by the EDPB (section 5.2.8).

5.2.1 Resolution 1973

Right from the earliest data protection law, reference was made to, and a special position was reserved for, sensitive data. Article 1 of Resolution 1973 (on the private sector) specified that the 'information stored should be accurate and should be kept up to date. In general, information relating to the intimate private life of persons or information which might lead to unfair discrimination should not be recorded or, if recorded, should not be disseminated.' It is interesting that these two seemingly unrelated elements are mentioned in one provision, for which no explanation is provided. The explanatory memorandum does provide as an example of data concerning a person's intimate private life information about her behaviour at home, her sexual life and her opinions and as an example of data that entails a risk for unfair discrimination data about a person's health and past criminal record. The provision basically

adopted the same structure as can still be witnessed in Article 9 GDPR, namely a prohibition of processing sensitive data, with exceptions, for example, as the memorandum provides, processing health data for counselling alcoholics or recording the political beliefs of members by political parties.

The provision was (much) stronger than the current regime on sensitive data in that it prohibited in full dissemination of those data. Disseminating was not understood in a limited fashion but as 'any transfer of information by a user to a third party, for example by a credit bureau to a bank.' It was also stronger in the scope of, what later became known as, sensitive data. The text refers both to data regarding 'intimate private life',[200] thus indirectly referencing Article 8 ECHR (the provision does not regard all data regarding private life, but only regarding intimate private life - where the boundary is drawn is not made explicit), and data which may lead to discrimination, thus indirectly referring to Article 14 ECHR. That article mentions discrimination on the basis of 'sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.' It is important to stress that not only does Article 14 ECHR contain a residual category and not only does the Memorandum to Resolution 1973 refer to 'etc.' when mentioning examples that could relate to a person's intimate private life, the fact that Article 1 of the Resolution holds a reference to both types of data equally means that it was not the attempt to define what are sensitive personal data exhaustively, but instead look at the (potential) effect of the data processing. The question of whether data could reveal parts of a person's intimate private life or be used for discriminatory practices was determinative of the status of the data.

5.2.2 Resolution 1974

Resolution 1974 (on the public sector) took a different approach. It also referred to both types of sensitive data, but did not contain a general prohibition, with exceptions, but only made special reference to these types of data when laying down the requirement of a legal basis, the purpose specification and the purpose limitation principle, holding that these principles must 'especially' be respected when these types of data are processed. In a way, this should be understood as applying normal public law requirements imposed on public sector organisations to the field of data processing. It is normal that public sector bodies only exert power on the basis of a law or similar regulation; it is normal that the powers that are transferred to them by the legislative branch are to be used for specific tasks only, and it is obvious that they, in principle, can only use the powers for those tasks that are provided in the law. The Memorandum also makes clear the article making special mention of sensitive data is basically a codification of the legality principle. Consequently, it is important to underline that the early data protection instruments were very strict on the processing of sensitive data by private sector organisations and very lenient towards public sector organisations doing so. This approach can still be witnessed in the many public interest grounds for legitimately processing sensitive data and the special rules contained in the Police Directive, though, over time, this sharp division has been toned down.

5.2.3 Convention 108

Convention 108 set out rules for both the public and the private sector and, in a way, applied the public sector regime of legality to both public and private sector organisations. It provides a prohibition of processing sensitive data, except when there is a national law, providing for appropriate safeguards. Though the explanatory report made clear that this requirement should not be limited to laws in the narrow sense, but also includes 'appropriate or specific regulations or administrative directives, as long as the necessary level of protection is secured',[201] it still imposed an important burden on private sector

[200] The term as such is used very seldom by the ECtHR. ECtHR, L. and V. v. Austria, appl. nos. 39392/98 & 39829/98, 09 January 2003. ECtHR, B.B. v. the UK, appl. no. 53760/00, 10 February 2004. ECtHR, Wolfmeyer v. Austria, appl. no. 5263/03, 26 May 2005. ECtHR, Big Brother Watch and others v. the United Kingdom, appl. nos. 58170/13 and 62322/14 and 24960/15, 25 May 2021. See also: ECtHR, Smith and Grady v. UK, appl. nos. 33985/96 and 33986/96, 27 September 1999. ECtHR, Lustig-Prean and Beckett v. UK, appl. nos. 31417/96 and 32377/96, 27 September 1999.
[201] Explanatory report, point 46.

organisations. It meant that a slightly more strict approach was taken than the Resolution from 1973 with respect to private sector bodies processing personal data because the Resolution left room for exceptions to the general prohibition, while the Convention requires a legal basis at all times; on the other hand, Resolution 1973 placed a strict prohibition on the dissemination of data to others, while the Convention leaves room for such practice when there is a legal basis.

The biggest change entailed the types or categories of data that were deemed sensitive. The Convention mentions data regarding racial origin, political opinions or religious or other beliefs, as well as personal data concerning health or sexual life and relating to criminal convictions. What is striking is that these categories neither include all previous elements mentioned with respect to intimate private life, e.g. data regarding home life is omitted, and the reference to 'opinions' is limited to specific opinions, nor includes all elements contained, e.g. in Article 14 ECHR, while it is formulated as an exhaustive list. In doing so, the Convention elevates the examples provided in the Resolutions to fixed and exhaustive categories of data.

However, the explanatory report denies such: 'The list of this article is not meant to be exhaustive. A Contracting State may, in conformity with Article 11, include in its domestic law other categories of sensitive data, the processing of which is prescribed or restricted. The degree of sensitivity of categories of data depends on the legal and sociological context of the country concerned. Information on trade union membership, for example may be considered to entail as such a privacy risk in one country, whereas in other countries, it is considered sensitive only in so far as it is closely connected with political or religious views.'[202] Thus, the Convention provides a list of data that are, in any case, to be regarded as sensitive, but allows countries to include additional categories dependent on their national context.

Though this is true, it is still important to note that the Convention does make a choice to emphasize the categories of data instead of the outcome of data processing, which had been the primary point of reference in the Resolutions. In addition, even if countries would adopt additional categories of sensitive data, this would most likely still result in a fixed and exhaustive list of data categories, as the explanatory report stresses that countries could adopt additional categories according to their national legal context, but makes no reference to adopting a residual category. This is implicitly confirmed by the explanatory report when it makes clear that the underlying goal of this provision was to prevent especially harmful practices from materialising, and that although such determination should normally be made in a context-dependent situation, there are exceptions. 'While the risk that data processing is harmful to persons generally depends not on the contents of the data but on the context in which they are used, there are exceptional cases where the processing of certain categories of data is as such likely to lead to encroachments on individual rights and interests.'[203]

5.2.4 Data Protection Directive

The same approach was taken in the EU Data Protection Directive 1995. As to the categories that were mentioned in the provision, the EU basically adopted the approach by the CoE, with small variations. The reference to trade union membership in the explanatory memorandum to Convention 108 was formalised, to racial origin was added ethnic origin, other beliefs (next to pollical opinions and religious beliefs) were made explicit as 'philosophical beliefs' and criminal data were mentioned separately. The Commission, in its proposal, underlined the same approach underlined in the explanatory report for Convention 108, namely that normally, the risks involved with data processing should be assessed in a context-dependent manner, but that for certain sensitive categories of data, it was clear that there are always risks involved.[204]

[202] Explanatory report, point 48.
[203] Explanatory report, point 43.
[204] Parliament suggested to also provide protection to 'or significant social circumstances including criminal convictions as well as any identification number issued by the public authorities' C 94 Volume 35 13 April 1992.

The novelty introduced by the EU Data Protection Directive is not that it was made clear that processing sensitive personal data should not only have a legal ground, but also serve an important public interest, or that criminal data were mentioned in a separate paragraph making clear that criminal convictions shall be held only in public sector files, but that it is made clear that sensitive data could also be processed on the basis of the data subject's consent.[205] Virtually no explanation was given concerning this introduction by the Commission in its proposal, and very little discussion exists on this point in the legislative process.

Still, it is important that Parliament suggested including an additional provision specifying: 'The Member States shall provide in their law for a ban on the processing of data of a strictly private nature in the private sector.'[206] This provision is interesting for several reasons. First, it introduces a reference to data of a strictly private nature, which seems an echo of the reference to data relating to intimate private life. Second, it again provides a ban, like the 1973 resolution, on the processing of sensitive data in the private sector, not even leaving an opening for processing such data without disseminating them. Third, the question arises of what would be the added value of the provision on consent, as processing sensitive data by private parties would be prohibited, and public sector organisations could only process sensitive data on the basis of a law.

Finally, there was a discussion in the legislative process of the DPD on both the question of whether the provision should also apply to the processing of sensitive data by non-for-profit organisations (as the initial proposal by the Commission included, besides the household exemption, a special status for processing of personal data by these types of organisations) and to the question of whether the processing also applied to the manual processing of sensitive data. The latter point is of special interest. The Commission's proposal for the Data Protection Directive applied to all processing of personal data, irrespective of whether such was done manually or automatically. Remarkably, it had proposed to apply the provision on sensitive data only to the automated processing of sensitive data. Why the provision on the processing of the most sensitive types of data should apply only to automated processing while the data protection regime, in general, would apply to all types of processing of personal data was left unexplained. Quite unsurprisingly, Parliament suggested that the provision applies to both manual and automated processing of personal data. This approach was adopted in the revised version of the proposal.[207]

Throughout the legislative process of the DPD, additional grounds for processing sensitive data were added. Starting with consent and a legal basis, a reference to non-for-profit organisations was included,[208] subsequently, situations in which interferences with human rights were unlikely were included in the list,[209] as was the employment context, the protection of the vital interests of the data subject or other persons and the establishment, exercise or defence of legal claims and data made manifestly public (presumably to be seen as the revised version of data processing which is unlikely to cause infringements in human rights).[210]

5.2.5 GDPR

Under the GDPR, both the definition of what qualifies as sensitive personal data and the grounds for legitimately processing those data have been revised. With respect to the definition, instead of 'processing of data concerning [] sex life', the GDPR refers to 'data concerning a natural person's sex life or sexual orientation shall be prohibited'. Thus, it includes data on the orientation itself, without

[205] COM(90) 314 final ~.sYN 287 and 288 Brussels, 13 September 1990.
[206] C 94 Volume 35 13 April 1992.
[207] C 311 Volume 35 27 November 1992.
[208] C 94 Volume 35 13 April 1992.
[209] C 311 Volume 35 27 November 1992.
[210] C 93 Volume 38 13 April 1995

actual information on sexual practices. The GDPR also adds new categories to the list: genetic data and biometric data for the purpose of uniquely identifying a natural person, next to the older category of data concerning health. These are also defined specifically in the GDPR. Genetic data refers to data concerning the inherited or acquired genetic characteristics of a natural person, which give unique information about the physiology or the health of that natural person and which result, in particular, from an analysis of a biological sample from the natural person in question. Biometric data concerns personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data. Data concerning health is also defined specifically under the GDPR, namely personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status.[211] In addition, the grounds for legitimately processing sensitive personal data has been expanded considerably.

It is relevant to summarise some of the main findings in the impact assessment, part of the legislative process of the GDPR.[212] That assessment made clear that under the Directive, several countries made use of the margin of appreciation left to the Member States to interpret and adapt the data protection framework to also provide protection to biometric data, genetic data, party membership, data from the judiciary, and even 'private life'. On the other hand, some national laws did not consider as sensitive data on ethnic origin, political opinions or philosophical beliefs. It suggested considering adding to the list of sensitive data, data relating to children, biometric data, genetic data and financial data. The inclusion of genetic data was thought to be in line with the jurisprudence of the European Court of Human Rights, in particular, the S and Marper v. UK case.[213] It also referred to a standing debate: does not every photo or video entail processing racial or ethnic data? The risk assessment did not give a clear-cut answer, but seemed to err on the safe side. 'Photos and images of persons, such as those published on the Internet or taken by traffic monitoring or other surveillance cameras, are especially problematic since they can reveal information about an individual's ethnic origin or health status.'

Though many changes have been made both to the categories of data qualified as sensitive and the grounds on which processing such data may be legitimate, and even more have been suggested during the legislative process of the GDPR, the biggest difference between the initially proposed text and the adopted version is that the initial texts gave the Commission the competence to set new rules with respect to processing sensitive data. 'The Commission shall be empowered to adopt delegated acts in accordance with Article 86 for the purpose of further specifying the criteria, conditions and appropriate safeguards for the processing of the special categories of personal data referred to in paragraph 1 and the exemptions laid down in paragraph 2.' This proposal, however, received much criticism, especially from parliament, and was finally changed to a provision which gives Member States competence in this field, albeit only with respect to certain types of sensitive data. 'Member States may maintain or introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or data concerning health.'

A final point of ongoing discussion, namely whether the processing of photo or video material per sé concerns the processing of sensitive personal data, was ended by recital 51: 'The processing of photographs should not systematically be considered to be the processing of special categories of personal data as they are covered by the definition of biometric data only when processed through a specific technical means allowing the unique identification or authentication of a natural person.'

5.2.6 Law Enforcement Directive

[211] GDPR, Article 4.
[212] Brussels, 25.1.2012 SEC(2012) 73 final.
[213] ECtHR, S. and Marper v. the UK, application nos. 30562/04 and 30566/04, 04 December 2008.

It is important that, in addition to the GDPR, there is also a so-called Police or Law Enforcement Directive,[214] which relates to the processing of criminal data by competent governmental organisations. Article 10 GDPR states: 'Personal data relating to criminal convictions and offences or related security measures may, pursuant to Article 6(1), only be processed under the control of official authority or if the processing is authorised by provisions laid down by Union law or Member State law which provide adequate safeguards for the rights and freedoms of the data subjects. Comprehensive records of criminal convictions may be kept only under the control of official authority'. In Data Protection Directive 1995, both articles (Art. 9 and 10 AVG) were contained in one provision. The processing of criminal data is a separate category - an even more 'special' special personal data - because, in general, the Police Directive applies to it and not the GDPR. Still, when non-law enforcement authorities process data that are considered criminal data, the GDPR applies.

5.2.7 CJEU

The Court of Justice has provided a broad interpretation of the grounds contained in the Directive. For example, in the *Lindqvist* case, a person had written on a blog that a colleague was on half-time on medical grounds because she had injured her foot. The question of whether having injured one's foot already qualifies as 'medical data' was only answered by the Court in a brief, staccato manner. 'In the light of the purpose of the directive, the expression data concerning health used in Article 8(1) thereof must be given a wide interpretation so as to include information concerning all aspects, both physical and mental, of the health of an individual. The answer to the fourth question must therefore be that reference to the fact that an individual has injured her foot and is on half-time on medical grounds constitutes personal data concerning health within the meaning of Article 8(1) of Directive 95/46.'[215]

Another case, that of *V.,* concerned the transfer of a medical file within the employment context.[216] The applicant complained about the sharing of her medical file within the employment context. The Tribunal held that, although the pre-recruitment examination serves the legitimate interest of the European Union institutions, which must be in a position to fulfil the tasks required of them, that interest does not justify carrying out a transfer of medical data from one institution to another without the consent of the person concerned. It pointed out that medical data are particularly sensitive data. Thus, it seemed to make a hierarchy between various categories of sensitive personal data, and seemed to attach to that fact the requirement of consent.

In his opinion, Advocate General Jääskinen, in the *Google Spain* case, stressed that search engines could not be regarded as controllers in relation to the personal data on source web pages hosted on third-party servers. Therefore, a reasonable interpretation of the Directive, according to the AG, required that the service provider is not generally considered as having that position. An opposite opinion would entail internet search engines being incompatible with EU law, a conclusion which the AG found absurd. Specifically, if internet search engine service providers were considered as controllers of the personal data on third-party source web pages and if on any of these pages there would be 'special categories of data', the activity of the internet search engine service provider would automatically become illegal, when the stringent conditions laid down in that article for the processing of such data were not met.'[217]

Interestingly, the Court observed in a slightly different tone that 'inasmuch as the activity of a search engine is therefore liable to affect significantly, and additionally compared with that of the publishers of websites, the fundamental rights to privacy and to the protection of personal data, the operator of the

[214] Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA.
[215] CJEU, C-101/01, *Criminal proceedings against Bodil Lindqvist* [2003] ECLI:EU:C:2003:596, para 50-51.
[216] CJEU, C- F-46/09, *V v European Parliament* [2011] ECLI:EU:F:2011:101.
[217] CJEU, C-131/12, *Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González* [2014] ECLI:EU:C:2013:424, para 89-90.

search engine as the person determining the purposes and means of that activity must ensure, within the framework of its responsibilities, powers and capabilities, that the activity meets the requirements of Directive 95/46 in order that the guarantees laid down by the directive may have full effect and that effective and complete protection of data subjects, in particular of their right to privacy, may actually be achieved'.[218]

But in the case of *GC a.o. v CNIL*, the Advocate General referred again to the words of the Advocate General in the Google Spain case. The AG in CNIL found that it was impossible to take an 'all or nothing' approach to the applicability of the data protection framework to search engines. He found that if the provision on sensitive personal data was applicable to search engines, the effect would be that any processing of the data listed in that provision would be prohibited, safe the handful of cases in which one or more of the exceptions would apply. A literal application of Article 8 would require a search engine to ascertain that a list of results displayed following a search carried out on the basis of the name of a natural person does not contain any link to internet pages comprising data covered by that provision, and to do so ex ante, and systematically, that is to say, even in the absence of a request for de-referencing from a data subject. This, the AG found neither possible nor desirable.

The AG quoted the judgment of the CJEU in the Google Spain case cited above and found that it is possible to draw two conclusions from that passage. Either meant that, in principle, every controller must satisfy all the requirements which that directive lays down, including those concerning sensitive data. Or, the limited responsibilities as a data controller of some parties, such as search engines, also has a bearing on the rules on processing sensitive data. The AG seemed to favour the latter approach and stressed that search engines cannot be expected to do ex ante checks on content referenced by them. 'The task of the operator of a search engine is, as its title indicates, to search, find, point to and make available by means of an algorithm that allows information to be found in the most effective manner. Conversely, it is not for the operator of a search engine to monitor, indeed to censure.'[219] Search engines can only be expected to do what they can practically be expected to do in relation to the rules concerning the processing of sensitive data, the AG found.

The Court, in its judgement, confirmed that approach and found that search engines were not exempted from respecting the rules on the processing of sensitive data, nor could they be expected to take on full responsibility. Their powers, capabilities and control over the data also had an effect on the responsibilities vis-à-vis the data protection framework. The rules on sensitive data can apply to the operator of a search engine only by reason of that referencing and thus via a verification, under the supervision of the competent national authorities, on the basis of a request by the data subject.

In the judgement of *La Quadrature du Net and Others*, the Court importantly made clear that even traffic and location data may reveal information on a significant number of aspects of the private life of the persons concerned, including sensitive information such as sexual orientation, political opinions, religious, philosophical, societal or other beliefs and state of health. It underlined that such data may allow very precise conclusions to be drawn concerning the private lives of the persons whose data has been retained, such as the habits of everyday life, permanent or temporary places of residence, daily or other movements, the activities carried out, the social relationships of those persons and the social environments frequented by them.[220]

Finally, in the case of *Latvijas Republikas Saeima,* the EU Court answered preliminary questions from the Latvian court concerning the relationship between the regimes at national level. The case concerns making information on 'penalty points' awarded for traffic offences available for re-use. With regard to the question of whether these data should be considered sensitive data, the Court stated that three criteria

[218] ECLI:EU:C:2014:317.
[219] ECLI:EU:C:2019:14, para. 51.
[220] CJEU, C-511/18, *La Quadrature du Net and Others v Premier ministre and Others* [2020] ECLI:EU:C:2020:791.

are relevant in order to assess whether the penalty points imposed for administrative traffic offences constitute an offence of a criminal nature (Art. 10 GDPR speaks of ('personal data relating to criminal convictions and offences or related security measures'): (1) the legal classification of the offence under national law, (2) the nature of the offence and (3) the severity of the sanction that can be imposed on the person concerned.

The first point, the Court recognised, in this case, should mean that the data are not categorised as sensitive data, because the national legal regime does not treat them as such. The second criterion concerns the question of whether the sanction in question primarily pursues a repressive objective. The measures in question are at least partly of a repressive nature, the Court found, but also pursue other objectives. As regards the third criterion, the Court observes that only road traffic offences of a certain degree of seriousness result in penalty points being awarded, that penalty points are generally imposed for such offences and that the accumulation of penalty points may itself have legal consequences, such as the obligation to pass a driving test or even to be banned from driving. Therefore, the Court considered that the data should be classified as sensitive data.[221]

It treated the case under the GDPR, not the LED. The Court ruled that in this case, the Police Directive does not apply, as the Road Safety Directorate, that processes the personal data, is not a 'competent authority' within the meaning of the Directive. Therefore, the GDPR applies. This means that the regime of the Police Directive is not applicable, nor is the heavy regime for special personal data, as contained in Article 9, but the 'ordinary' regime for 'ordinary' personal data, as contained in Article 6 GDPR. However, the Court stresses that the additional requirements as mentioned in Article 10 GDPR apply, namely that the processing must be under the supervision of a government organisation and that additional protective measures have been taken.

5.2.8 EDPB

Lastly, the EDPB discussed the use of sensitive data for political campaigns. On the one hand, it is argued that inferring the probability that a data subject will vote for a certain party based on monitoring its activity, *e.g.,* visits to web pages with political ideology content would constitute a special category of data processing. On the other hand, it is also recognized that the processing of a 'mere statement or a single piece of location data or similar' revealing that a user visited (once or several times) a place typically visited by persons with certain religious beliefs will generally not constitute sensitive data processing.[222]

5.3 Technical developments

This section will provide insights gained on the technologies that can be used for pseudonymising and de-pseudonymising and encryption and decryption gained through the literature study (section 5.3.1), the interviews conducted for this study (section 5.3.2) and a workshop held for this study (section 5.3.3).

5.3.1 Literature study

This section will briefly shed light on three important studies that problematise the categories of sensitive data and the distinction between sensitive and non-personal data.

First, Cabañas et. al conducted a scientific study with the aim of shedding light on the surreptitious inference of sensitive data categories for advertising purposes from the activity of social network users.[223] According to the authors, certain social networks' ad preferences assign to each one of their

[221] CJEU, C-439/19 - *Latvijas Republikas Saeima* [2021] ECLI:EU:C:2021:504.
[222] EDPB, 'Statement 2/2019 on the use of personal data in the course of political campaigns', 13 March 2019.
[223] Cabañas, J. G., Cuevas, Á., & Cuevas, R. (2018). Facebook use of sensitive data for advertising in Europe. arXiv preprint arXiv:1802.05030.

users specific parameters derived from their personal data, ordinary or sensitive. These parameters are usually assigned based on the user's activity in the social network itself or through other means, such as external websites, apps, or online services, and can be exploited by advertisers to define the target audience of the advertising campaign. Whereas predefined parameters based on 'ordinary' personal data provided by the data subject are presented by design, such as location, gender, age, or language, it is also possible to leverage an 'interest parameter' for highly customized targeting. By entering any free text in the interest parameter, the advertising preference system will suggest parameters linked to that text to achieve a very detailed audience. As confirmed by the authors, the introduction of indirect identifiers, such as 'Islam', 'homosexuality' or 'reproductive health', matched predefined parameters used by the social network while constituting special categories of data. In this way, the authors demonstrated that such sensitive parameters had already been assigned to the users of the social network, which evidences the systematic sensitive categorization of user's activity. While the study was not intended to investigate the parameter assignment process, questions arise concerning the legal qualification of the data on which such sensitive inferences are based.

The results of the study indicate that 2092 ad preferences out of the 126.000 analysed constituted sensitive data. In other words, 1,66% of the analysed ad preferences constituted sensitive data. Whereas the absolute amount of sensitive data tags remained limited in number, the authors further demonstrated that more than 73% of the social network users were labelled with, at least, one of the top 500 most popular sensitive ad preferences found. By extrapolation, this implied that 40% of the EU citizens using the social network were labelled with a sensitive data parameter. Whereas the details of the conversion process cannot be thoroughly assessed, the results of the study seem to indicate that data conversion from ordinary provided and observed data to sensitive data may be possible where social media platforms carry out user profiling for sensitive purposes.

Second, in another experiment on social media, inferences of policy positions based on users' connections were carried out by Barberá.[224] For these purposes, social network users were scaled along their common ideological dimension based on the individuals they 'followed'. The study rested on the assumption that the analysed social network was homophilic. In other words, it was assumed that ordinary users of the social network and their political counterparts interacted within the same symbolic framework. The author justified the assumption based on two reasons. First, according to previous sociological research,[225] the author argued that social network users follow other accounts whose ideology is similar to theirs and tend to interact and relate more often with users who exhibit similar traits. Second, the author assumed that this homophilic pattern was strengthened by 'selective exposure'.[226] Selective exposure is a documented phenomenon[227] whereby users show preferences for 'opinion-reinforcing political information' that is aligned with their ideology. In light of the above considerations, Barberá was able to infer valid political positions for common users and political actors from the structure of the 'following' links between these two groups of users. To do so, the author developed a Bayesian model that estimated the ideal points for a large sample of users in different countries, including the United States, the United Kingdom, Germany, Spain, Italy and the Netherlands. Individual estimates were derived from the observed 'following' decisions of the users, under the assumption that their behaviour was instrumentally rational. The resulting point estimates of political actors were further compared with their respective roll-call votes for confirmation purposes, whereas the point estimates of ordinary users were compared with their political position published in their profile descriptions. Accordingly, individuals who identified themselves as 'liberal', 'moderate', and 'conservative' on their profiles were successfully characterized on the 'left', 'centre', and 'right' parameters of the resulting ideological scale. Further validation of the results was introduced by

[224] Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. Political analysis, 23(1), 76-91.
[225] McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. Annual review of sociology, 27(1), 415-444.
[226] Bryant, J., & Miron, D. (2004). Theory and research in mass communication. Journal of communication.
[227] Lazarsfeld, P.F., Berelson, B., and Gaudet, H. (1944). 'The people's choice: How the voter makes up his mind in a presidential election' New York: Duell, Sloan and Pearce.

matching a sample of the analysed accounts with campaign contribution records and with voter registration files, as well as with manifestos and surveys. The results of the study call, again, for unexplored inference potential of social network behaviour.

Similar to the previous study of Cabañas et. al, the study conducted by Barberá also points out a hasty conversion of 'ordinary' personal data into sensitive data by way of inferences aimed at generating 'ideology estimates that could prove useful in political science'. However, in contrast to the previous study, some nuances should be noted. First, the processing of personal data that may reveal a political orientation when the user has explicitly stated this in his or her social network profile should, in the opinion of the EDPB, constitute sensitive data processing per se.[228] Therefore, if, for instance, a user explicitly states in his or her individual profile that he or she is a member of a certain political party, the processing of that data should amount to sensitive data processing. This would require the data controller to rely on an exemption in Art. 9(2) GDPR cumulatively with a legitimate legal basis under Art. 6 GDPR. Such an instance may lead, in certain contexts, to the coexistence of different data protection regimes for users who explicitly declare their political ideology in their profiles and those who do not. Of course, in these cases, the purposes of the processing must also be taken into account in ligh of the prominent role that purpose specification plays in the determination of the legal nature of the processing. The EDPB illustrates this by means of the following example: where a social media provider uses provided ordinary personal data, such as ‘age, interests, and address’ and ‘combines it with observed data about website visits and “likes” on the social platform’ (emphasis added) to infer the political ideology of the data subject for policy categorization purposes, special categories of data processing should amount. In other words, the purpose of political categorization is also determining factor for the legal qualification of personal data processing. Even if no explicit mention of the political ideology is being made by the user, the processing of said data would be considered sensitive data processing where the purpose followed by the controller is political categorization. The same situation may occur where 'ordinary' personal data, such as 'following' links, are processed for political categorization purposes. In these cases, conversion from 'ordinary' personal data to sensitive data would also be at stake. Contrarily, when large amounts of personal data of a potentially sensitive nature are processed, including 'following' links, said processing may not automatically account for a special category of data processing as long as, having considered appropriate measures to prevent inference or targeting, it does not result in inferences of special categories of data.

Third, further insights can be obtained from the conversion of ordinary personal data into sensitive data in areas outside social media. In the health research field, for instance, Allerhand et al. developed a methodology for diagnosing Parkinson’s disease from the interaction of users with a search engine.[229] By means of tracking software that collects the position of the user's mouse cursor on the computer or browser page, the authors developed a system to diagnose the early stages of Parkinson's disease. The system relied on the detection of abnormalities in the motor behavior of the user through spontaneous interaction with the search engine, including the rigidness and shakiness of the moves, as indicators of primary symptoms of the neurodegenerative disease. They used an unsupervised representation learning technique to predict event-level mouse movements and extract features for the diagnosis model. According to their results, they were able to achieve a true positive rate of 0.92 or, in other words, to satisfactorily predict Parkinson's disease with an accuracy of 92% by employing all features used for prediction in the experiment and aggregating all interaction sessions of the user.

While the results seem to clearly indicate that mouse tracking data can help in detecting users at the early stages of Parkinson's disease, data protection concerns as to the legal categorization of mouse tracking data can be more disputable. Most importantly, the easiness of deriving sensitive information from alleged ordinary personal data, such as mouse movements, triggers again the regulatory question

[228] EDPB, 'Guidelines 8/2020 on the targeting of social media users', 13 April 2021.
[229] Allerhand, L., Youngmann, B., Yom-Tov, E., & Arkadir, D. (2018, October). Detecting Parkinson's Disease from interactions with a search engine: Is expert knowledge sufficient?. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management (pp. 1539-1542).

of whether these types of personal data should methodically constitute special categories of data or, more specifically, data concerning health. Given the broad definition of data concerning health, one could argue that, in light of the previous experiment, provided the necessary processing means, any mouse movement could have the 'potential' to inherently relate to the physical or mental health of a natural person which reveals information about his or her health status, and therefore constitute data concerning health. However, this would consequently lead to a systematic categorization of mouse movements as special categories of data, thus requiring search engines or webpages tracking the cursor movements of their users to correspondingly comply with the stricter data protection regime for sensitive data. While such an assumption may be legitimate in certain cases, for instance, where the purpose of the processing is the determination of the health status of the person, it is legally doubtful whether ordinary cursor movements arising from the daily interaction of users with the Internet may sufficiently trigger stricter controller's obligations.

5.3.2 Results from interviews

This section will discuss some of the main findings gained through the interviews conducted for this study. The full interview reports may be found in the annex to this report.

From the interviews, the following insights were drawn with respect to sensitive personal data:

- Scope of sensitive data: experts question the fixed categories of sensitive data used in the GDPR. For example, the financial position, socio-economic background or income of a person could be treated as sensitive data.
- Cultural aspect: what is or should be considered sensitive personal data varies per region or country, which is why working with one fixed list of types of sensitive data for all EU countries alike may be challenging.
- Contextuality of sensitive data: what is or is not sensitive, at least from a technological perspective, does not depend on fixed types of data. Data processing can be sensitive and harmful even without the categories of data listed in the GDPR or can be non-harmful even if one or more of the types of data categorised as sensitive are processed.

5.3.3 Results from workshop

The workshop held for this study yielded the following results:

- Categories of sensitive data: the current categories are perhaps not necessarily the most sensitive ones. Financial information, location data, poverty, metadata and so forth could also be included in the list.
- Exhaustive list: a number of experts suggested that the list contained in the GDPR should not be exhaustive.
- Holistic approach: a possible alternative could be regulation without categories of sensitive data but looking at the sensitivity of the processing as a whole and having the levels of requirements and obligations put on data controllers depending on the sensitivity of the data processing operation as a whole.

5.4 Analysis

Four main tensions between the legal and the technical realm have emerged from this chapter:

1. Like the distinction between personal and non-personal data, the legal regime makes a binary distinction between non-sensitive personal data and sensitive personal data. Yet, there has been a shift in the legal definition of sensitive data. Initially, data were considered sensitive when processing could have a significant impact on the private life of a data subject or entailed a significant risk for discriminatory practices. Open-ended and non-exhaustive examples of data

categories were provided. Over time, the legal regime has shifted towards defining concrete and exhaustive lists of types of data that are considered sensitive, despite the context or the processing operation concerned. Technical experts challenge the binary distinction between sensitive and non-sensitive personal data. They often rather approach each data processing operation on a case-by-case basis, taking a holistic understanding of the potential risk, the harm entailed when the risk materialises, and the possibilities to achieve the goals without the data concerned. On the basis of that assessment, the level of risk and sensitivity is determined, as well as the level of protection and security that is needed.

2. This also means, on a more abstract level, that from a technological perspective, it is not the data as such that is determinative of the sensitivity of the data processing operation, but (also) other aspects, such as the technologies used, the amount of data, the goal of the data processing operation and the application of the data processing operation.
3. If the current approach of providing an exhaustive list of sensitive personal data should be maintained, technical experts suggest several additional categories should be included, such as financial information, location data and metadata.
4. Like with the legal distinction between non-personal and personal data, technical experts point to the fact that sensitive information can often be derived from non-sensitive personal information. Thus, although the legal regime makes a binary division between the two, in reality, the lines are more fluid.
5. Finally, what complicates matters is that there is no uniformity from a legal perspective on the matter of inferences of sensitive personal information from personal or non-personal data. The former Article 29 Working Party,[230] the European Data Protection Supervisor,[231] the European Data Protection Board,[232] and researchers[233] have all concerned themselves with the question of whether inferences of personal data still constitute personal data. For sensitive data, inferring information is a critical issue. If sensitive information can be inferred from non-sensitive personal data, this means that it is more difficult to work with a legal binary approach to sensitive and non-sensitive data. Article 29 Working Party has pronounced itself about this issue by differentiating between 'provided' and 'observed' data and 'derived' and 'inferred' data.[234] While provided and observed data refer to data actively and knowingly provided by the data subject or observed from the activity of the data subject at the event level, derived and inferred data refer to data that is created by the controller on the basis of provided or observed data. According to an early opinion of Article 29 Working Party, only personal provided or observed data form part of the right to portability, while derived or inferred data 'will typically not' fall within the scope of the right to data portability. Although no explicit pronouncement was made on the nature of derived or inferred data, the fact that such kind of derivation or inferences were not included in the right to data portability implies a distancing effect toward the controller's personal data protection obligations. This line of thought appears to be, however, reversed in a later opinion on automated individual decision-making and profiling,[235] where profiling is defined as the process of 'creating derived or inferred data about individuals' which constitutes '*new personal data* that has not been provided directly by the data subjects themselves' (emphasis added). In this line, Article 29 Working Party defends the conversion of ordinary personal data into sensitive data where profiling 'create[s] special categor[ies] [of] data by inference from data which is not special category data in its own right but becomes so when combined with other data'. Hence, it appears that Article 29 Working Party settled the matter of what the creation of

[230] Article 29 Working Party, 'Opinion 03/2013 on Purpose Limitation', 2 April 2013.
[231] EDPS, 'Opinion on Online Manipulation and Personal Data', 19 March 2018.
[232] EDPB, 'Guidelines 8/2020 on the targeting of social media users', 13 April 2021.
[233] Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. Colum. Bus. L. Rev., 494.
[234] Article 29 Working Party, 'Guidelines on the Right to Data Portability', 13 December 2016 As last Revised and adopted on 5 April 2017.
[235] Article 29 Working Party, 'Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679', 3 October 2017 As last Revised and Adopted on 6 February 2018.

derived or inferred data constitutes in its more recent pronouncements, but discussions over this matter are still ongoing.

Chapter 6: Analysis

6.1 Introduction

This final chapter will summarize the main conclusions of this study (section 6.2). It will discuss the regulatory objective of the privacy and data protection regime (section 6.3). Subsequently, it will suggest that determining the bottlenecks and the dangers of over- and under-regulation depends on the discussions as outlined in section 6.3 as well as others (section 6.4). Finally, it will provide an overview of regulatory alternatives to solve regulatory gaps that have been suggested in literature and elsewhere (section 6.5), sketch five hypothetical scenarios for the regulatory approach to the data landscape (section 6.6) and answer the research questions for this study (section 6.7).

6.2 Summary of main findings

In essence, this report dealt with the tension between two regulatory approaches: a contextual one and a categorical one, an approach that takes into account the circumstances of the case and an approach that is based on fixed definitions and clear regulatory rules attached to those definitions. Each of these approaches has clear benefits and disadvantages. The first one is able to take into account all relevant aspects per scenario, is more adaptive to changing circumstances, and so does not run the risk of becoming outdated or being circumvented. However, fluid and contextual regulatory approaches have the disadvantage that they are vague and provide little legal certainty, both to the data controller and to the data subject. The second approach solves this problem: it gives a clear set of definitions and categories and attaches to those a clear set of rules. Yet the disadvantage is also clear: it runs the risk of being circumvented, becoming outdated and is less granular than a contextual approach.

This research has shown the deep ambivalence that runs through the regulatory approach to data protection on this point.

One first sight, the categorical approach is most apparent. It was shown that the disconnection of the right to data protection from the right to privacy had to do with a decontextualization of the right. In the human rights framework, a claim is assessed on both the *ratione materiae* (does the matter complained of fall under the material scope of the article invoked?) and the *ratione personae* principle (can the applicant claim to be a victim?). With respect to that second principle, there is a significant threshold, as applicants must be able to show that they have suffered from direct, individualizable and substantial harm. Under the data protection framework, both principles are merged. This means that any processing of personal data, however mundane and small, is considered personal data processing, to which the GDPR applies. This means that a contextual or harm-based element that is essential to evaluations of human rights is removed under the data protection regime. The application of the data protection regime, different from, for example, the right to privacy, does not depend on the question of whether there has been harm inflicted on a claimant or rights bearer.

In addition, it is clear that the data protection framework works with a clear and binary distinction between personal and non-personal data. This study showed that the EU has provided personal data with the highest form of legal protection in the world through the GDPR and the LED, while with respect to processing non-personal data, it explicitly discourages restrictions set by private and public sector organisations. Because the distinction between personal and non-personal data is a binary one, the question of whether a dataset is categorised as either one will mean a regulatory difference of 180 degrees, though the proposed Data Governance Act may complicate matters.

A binary approach can be witnessed with respect to both pseudonymous data and sensitive personal data: data are either pseudonymous or they are not, and personal data are either sensitive or they are not. With respect to the latter type of data, the categorical approach is even more apparent. The GDPR sets a limited and exhaustive list of types of data that are considered sensitive, namely: data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation. The processing of such data is, in principle, prohibited.

Though largely falling outside the scope of this research, similar binary approach is prevalent with respect to the difference between content communications data, between that what a person says over the phone, in an e-mail, in a letter or through another communicational method, and metadata, which are data about the communicational activity itself, such as who communicates with whom, how often they communicate, where they are located, what type of communicational technique is used, etc. In the EU, locational and traffic data, on the one hand, and content communications data, on the other hand, are regulated differently in the e-Privacy Directive and the e-Privacy Regulation, which is now under discussion. The latter will reserve a special position for the separately defined category of metadata. Under the European Convention on Human Rights, content communications data fall per se under the right to informational secrecy (the fact that private communication is monitored is a harm per sé), while metadata is, in principle, not covered by the right to private communication.

A final point that should be underlined is that the data protection framework as a whole is based on binary distinctions and is marked by a categorical approach. For example, it sets out clear differences between various actors, such as the data controller, the data processor and the data subject. Each of those actors has a clearly defined role, a set of obligations and rights and regulatory responsibilities. A party cannot at the same time be a data processor and a data controller with respect to the same data processing operation: it is either or. Similarly, with a data processing operation, a party is either a (joint) controller or a data subject.

On the other hand, a contextual approach is visible. For example, though the distinction between personal and non-personal data is binary, the definition of personal data includes a contextual aspect. The notion of ‘identifiable’ means that data that at this moment in time does not allow for the identification of an individual, but in the future will, are already to be considered personal data now. This requires an assessment of the likely future status and use of data. Similarly, recital 26 GDPR includes a high number of contextual elements for determining whether data are personal or not: ‘To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.’

Furthermore, although the category of pseudonymous data is in itself binary – data are either pseudonymous or they are not – it is seen by many as an intermediate category between personal and non-personal data. Pseudonymous data are not anonymous and thus the GDPR applies, yet they are not so easy to connect to an identified individual as non-pseudonymised personal data. That is why the GDPR allows for a number of exceptions to the rules and obligations it lays down when data are pseudonymised. Similarly, although the distinction between non-sensitive and sensitive data is often presented as absolute, all the different rights and obligations apply to both sensitive and non-sensitive personal data alike. The only difference is the legitimate ground for processing the data (Article 6 and Article 9). Although Article 9 takes as starting point that processing sensitive data is prohibited, it lists a high number of exceptions to this prohibition, making the difference between the processing of

sensitive and non-sensitive data practically redundant. In fact, these exceptions are closely similar to the processing grounds mentioned in Article 6, except for ground 6(1)(f).

It should also be underlined that when the data protection regime applies, most obligations and requirements are actually context-dependent, meaning that in general, the more data are gathered, the more sensitive those data are, and the higher the risk entailed with data processing, the more parties involved, etc., the stricter the rules and obligations apply. This contextual approach applies to the obligation to implement a data protection policy, adopt technical and organisational security measures and embed data protection choices by default and by design in the infrastructure of an organisation. The documentation requirement does not apply to small organisations that do not engage in risky processing operations, a data protection impact assessment only needs to be executed where potential harm is likely, and private sector organisations only need to appoint a data protection officer when their core activities consist of regular and systematic monitoring of data subjects on a large scale or large scale processing of sensitive data and the data breach notification depends on the harm likely resulting from the breach. Consequently, the core of the data protection framework is highly contextual.

In addition, both courts went beyond the fixed categories as entailed in the legal regime. For example, the CJEU has made clear that there are differences between the types of sensitive data, stressing that medical data is especially sensitive. In addition, the ECtHR has made clear that metadata does not per se deserve a lower level of protection than content communication data, but that much depends on the context. It provides protection to metadata both if they are linked to the content of communications and when processing metadata has an impact on citizens' private life. Similarly, advisory bodies such as the former Article 29 Working Party and the current European Data Protection Board have taken a contextual approach to the various data distinctions and definitions.

Finally, it has to be underlined that although the European approach to privacy and data protection is often contrasted with the American one, the former adopting an omnibus approach, the latter taking a sector approach, the contrast is less sharp than often imagined. It is clear that the EU explicitly does separate two contexts when it separates the law enforcement context, to which the LED applies, from other contexts covered by the GDPR. In addition, the GDPR promotes the use of codes of conduct, through which sectors can adopt their own interpretation and specification of the rules as entailed in the general data protection regime. The fact that this possibility has not gained ground, inter alia, because sectors fear the administrative burden of performing oversight and handling complaints does not mean that it is not possible to take a sectoral and, thus, more contextual approach to the data protection regime.

From the technological perspective, an equally ambiguous picture emerges from this study.

On the one hand, technological experts, question the categories as defined in law. To them, it is often unclear where the boundary lies between non-personal and personal data, between pseudonymous and anonymous data and between sensitive and non-sensitive data. The legal definitions are regarded as too vague and complex to provide clear guidance for decision-making in practice. In addition, the logic behind the data categories as defined in the regulatory regime is questioned. For example, why is pseudonymous data granted a special status vis-à-vis other privacy-preserving technologies? Also, the list of categories of sensitive data is seen as too limited. Suggestions have been made, inter alia, to also include in the list financial data, data about minors and locational data.

Technologists themselves often propagate a contextual and holistic approach. They see little value in setting absolute categories, but rather determine the level of protection and precautionary measures that need to be taken in light of the potential risk and harm that the processing operation entails. For example, it is not seen as determinative that medical data are processed when determining which protective regime to apply. Some processing operations concerning medical data may indeed be very sensitive, while others may entail no significant risk. Some processing operations not concerning sensitive data,

however, can be highly risky and potentially harmful. This means that the level of privacy and data protection and the types of technologies used to ensure technical and organisational security depends on the context and on a case-by-case evaluation of the risk. This also applies to, for example, the type of pseudonymisation or anonymisation technique deployed. The type of context and processing operation dictates which technique is most suitable.

Perhaps more fundamental, technological experts question whether the rationale underlying the separate data distinctions and definitions is still viable. It is easy, for example, to derive content communication data from metadata, just like sensitive personal data may be inferred from non-sensitive personal data or even non-personal data. Data categorisations work well in a world in which the data are relatively stable and fit into one regime or another, while in reality, datasets are in flux and change constantly. Thus, it can be questioned if it even is possible to work with data categories at all. In addition, they stress that processing large quantities of non-personal data and making decisions on the basis of aggregated data, group profiles, and longitudinal patterns may be just as or even more intrusive, harmful and impactful as processing personal data; processing personal data, such as financial data or data about minors, may be more impactful than processing established categories of sensitive data; the bulk collection and analysis of metadata may have a bigger impact than the collection of content communications data; etc.

On the other hand, the contextual elements in the regulatory regime are approached with caution. The contextual elements in the definition of personal data and the description of anonymisation, for example, are regarded as imprecise and abstract. In addition, experts ask for more criteria or factors from the regulator that enable them to determine whether a dataset falls into the category of non-personal, pseudonymous, personal or sensitive personal data. Many also believe that partial technological neutrality (it is not technological neutral, inter alia, because it explicitly mentions pseudonymous data and because it distinguishes between automated processing of data and unstructured manual processing) is unhelpful. There needs to be more clarity as to what level of anonymity, what level of pseudonymisation, etc., is to be attained for a dataset to qualify in a certain category and how to attain that. Which pseudonymisation techniques, for example, are allowed and which ones are not?

Also, it is difficult for many non-lawyers that there are so many legal regimes and rationales applicable. In the EU, there is the GDPR, the LED, the Data Governance Act, the Data Act, the AI Act, the Digital Markets Act, the Digital Services Act, the Open Data Directive, the e-Privacy Regulation, etc. Each of these has its own set of rules, defines its own set of actors, and distinguishes between its own set of data categories. Where legal rationales conflict, for example, the rationale of data protection in the GDPR and the rationale of openness and re-use in the Open Data Directive, the EU does not provide any guidance on how to reconcile those rationales in specific situations.

Regulatory alternatives that were identified through the literature study, the interviews and the workshop can be roughly divided into three approaches.

The first approach is set on providing clarifications of the current regime or suggesting small alternations while leaving the general regulatory system intact. Such regulatory suggestions include providing more clarity on the boundaries between the various data concepts, determining the techniques that are deemed appropriate in certain contexts and setting out best practices. In addition, alterations could include, for example, expanding the list of sensitive personal data.

A second approach is to adopt a more contextual approach. This could mean, for example, letting go of the category of sensitive personal data in the GDPR and instead assessing each processing operation on a case-by-case basis, determining in a holistic manner the potential harms and risks involved and the level or protection required. In a similar vein, a more gradual approach to the distinction between personal data and non-personal data has been advocated.

A third approach is to strengthen the categorical approach adopted in privacy and data protection law, but to alter the categories and the regulation thereof. For example, it has been suggested that a strong regulatory regime should be in place for the collection and analysis of metadata. This regime should not be similar to the regime in place for the collection and analysis of content communication, but instead, take account of the distinctly different nature of the processing of metadata. Another example is the processing of non-personal data, to which a GDPR light regime may be applied, ensuring that such processing of such data needs to abide by a minimum set of due process requirements.

6.3 Regulatory objective of data protection law

In order to assess whether there are regulatory gaps and, if so, where they lie and which regulatory alternative is most suitable, it is necessary to assess what the regulatory objective of the privacy and data protection regime is. This is a matter of debate.

In general, there exist two ideal-type models of data regulation. The first one is to give control to individuals over their personal data. In its most extreme form, scholars have suggested that if individuals would gain property or other control rights over their data,[236] they would be able to adequately represent and protect their own interests against the multinationals and governmental organizations that intend to use their data.[237] This model has clear advantages as it grants citizens autonomy over their personal data and steers away from any form of paternalism. In addition, it sets no absolute boundaries on what organisations can and cannot do with personal information, but connects that question to each individual's preferences, which may vary significantly per person.[238] To link it to the discussion of the previous section, it is a highly contextual form of regulation.

But this model also has clear disadvantages. The capacity of citizens to make choices according to their best interests is limited in practice both because of the complexity of most contemporary data-driven processes involving biometric data, artificial intelligence and profiling because of the multitude of processes which contain the data of an average citizen, and because of the information-asymmetry between data-driven organisations and the average citizen.[239] In addition, many of the data-driven processes affect large groups in society or the population in general; leaving it to each and every individual citizen to assess such processes and their potential flaws individually would mean a privatisation of structural problems and would result in well-educated citizens protecting their personal data better than would already marginalised groups.[240] In addition, it misses out on the protection of societal interests that transcend that of individuals.

A second model is to rely on legal standards and governmental enforcement of those standards. Just like there are minimum safety requirements for cars – a citizen can simply not legitimately buy a car that does not meet the legal safety standards – there are minimum requirements for legitimately processing personal data. It is not left to citizens to assess whether these rules are met, but to an independent governmental organisation, which has the authority to both investigate data-driven organisations and set sanctions and fines when they violate the rules. This means that legal protection is provided to citizens, without them having to assess the validity, legality and desirability of each individual data process that contains her data on her own.

However, this model too has its particular disadvantages. Citizens may be limited in having their data processed against their will, and legal standards are often too general, absolute and inflexible and easily

[236] Samuelson, P. (2000). Privacy as intellectual property?. Stanford law review, 1125-1173.
[237] Mun, M., Hao, S., Mishra, N., Shilton, K., Burke, J., Estrin, D., ... & Govindan, R. (2010). Personal data vaults: a locus of control for personal data streams. Proceedings of the 6th International Conference, 1-12.
[238] Lazaro, C., & Metayer, D. L. (2015). Control over personal data: true remedy or fairy tale. SCRIPTed, 12, 3.
[239] Cate, F. H., & Mayer-Schönberger, V. (2013). Notice and consent in a world of Big Data. International Data Privacy Law, 3(2), 67-73.
[240] Lanzing, M. (2016). The transparent self. Ethics and Information Technology, 18(1), 9-16.

become outdated in the constantly developing data-driven environment.[241] In addition, it is practically impossible for one governmental organisation to assess all data processing operations[242] and difficult to ensure that parties based in other territories adhere to national standards. This means that supervisory organisations, such as Privacy Commissioners and the Data Protection Authorities in Europe, usually only focus on the bigger data processing operations that have the biggest potential impact.

Both approaches are visible in the GDPR, though it is clear that the latter approach is dominant.

Both privacy and data protection are considered independent fundamental rights under the EU Charter of Fundamental Rights. Article 7 provides: 'Everyone has the right to respect for his or her private and family life, home and communications.' Article 8 holds: '1. Everyone has the right to the protection of personal data concerning him or her. 2. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified. 3. Compliance with these rules shall be subject to control by an independent authority.' There is a discussion on the right interpretation of both of them, which boils down to the question: should the right be seen as a prohibition on the one interfering with the human right, or should the human right be seen as a right to control?

With respect to the right to data protection, reference is made to Article 5 of the General Data Protection Regulation, which is seen as the backbone of the law. It holds that personal data should be processed lawfully, fairly, and in a transparent manner, collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes, adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed, accurate and, where necessary, kept up to date, kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed and processed in a manner that ensures appropriate security of the personal data. These are all obligations posed on the data controller that are applicable independent of any rights being invoked by them.[243]

On the other hand, there are increasingly many rights attributed to data subjects in the data protection regimes, the GDPR includes the right to access, the right to copy, the right to information, the right to object, the right to erasure, the right to rectification, the right to data portability, the right to restrict, the right not to be subject to automated decision-making and the right to file a complaint. In Europe, especially due to German influence, the notion of informational self-determination has become increasingly popular.[244] Thus, some argue that, rather than the obligations posed on data controllers, the rights of data subjects are the core of the data protection regime.

The same discussion plays a role in the human rights framework in general and the right to privacy in particular. The European Convention on Human Rights, in Article 8, provides protection for the right to privacy. Initially, citizens were not allowed to submit a complaint to the European Court of Human Rights themselves.[245] Member States could submit inter-state complaints, or a Member State or the European Commission on Human Rights could send an 'individual submission' to the Court when they were convinced that that claim had a broader significance, transcending the particularities of that specific matter.[246] It was believed that the majority of the cases would be inter-state complaints, and of the individual cases, many would be brought by legal persons, for example, civil society organisations

[241] Zarsky, T. Z. (2016). Incompatible: the GDPR in the age of big data. Seton Hall L. Rev., 47, 995.
[242] Bennett, C. J. (2018). Regulating privacy. Cornell University Press.
[243] Mahieu, R. (2021). "The Right of Access to Personal Data: A Genealogy". Technology and Regulation 2021 (August), 62-75.
[244] Hornung, G., & Schnabel, C. (2009). Data protection in Germany I: The population census decision and the right to informational self-determination. Computer Law & Security Review, 25(1), 84-88.
[245] ECHR 1950, Article 48.
[246] Van der Sloot, B. (2017). Privacy as virtue (Cambridge: Intersentia).

and groups. Inter-state complaints per sé do not revolve around harm claimed by the applicant but concern a general policy or legal system that is deemed to be in violation of the Convention.[247]

Adopted in the wake of the Second World War, the Convention was intended to address larger societal concerns over the abuse of governmental power by totalitarian regimes.[248] Not the individual harmed through a specific action by the executive branch, e.g. a governmental official unlawfully entering a home or wiretapping a telephone, but the stigmatisation of minorities, Stasi-like governmental surveillance and anti-democratic practices were top of mind.[249] Hence, emphasis was placed on negative obligations for states, that is, not to abuse their powers, and negative rights for citizens, that is, not to be interfered with their rights, rather than on subjective claim rights for natural persons to protect their personal interests in concrete cases and on positive obligations for states to pursue their desired life path.[250]

Over time, however, the Convention structure has changed. Natural persons have been allowed direct access to the Court, the possibility of inter-state complaints never gained ground[251] , and the Court has barred groups from submitting a claim as a group[252] and has been hesitant to allow legal persons to invoke the right to privacy, as it feels that this doctrine primarily provides protection to individual interests and not to societal ones.[253] In addition, the Court has adopted a very tight approach to assessing individual claims: a natural person needs to demonstrate concrete, substantial and individualizable harm that has already materialised and bears a causal relation to the matter complained of.[254] Furthermore, the Court has chosen to take a case-by-case approach, therewith choosing to provide a solution in the concrete circumstances of the case rather than focussing on general legal questions that have significance for other cases or society as a whole.

Hence, yet again, two interpretations of the right to privacy exist: there is the group that primarily sees the right to privacy as a subjective claim right attributed to natural persons to protect their private interests, and there is the group that sees human rights, including the right to privacy, as primarily putting an obligation on states not to abuse their power and setting out limits and conditions for the use of power. In addition, there is the contextual case-by-case analysis that the European Court of Human Rights commonly takes, and there is the approach in which the Court assesses on a general and abstract level the validity of legal regimes and policies as such, an approach which was propagated when the Convention was adopted and has recently been reintroduced in the jurisprudence of the ECtHR when dealing with the bulk collection of metadata by intelligence agencies.

What makes the assessment even more complex is the fact that the data protection regime does not only have a protective objective, but also acknowledges as an objective facilitating the processing of personal data in the EU. Article 1 GDPR specifies: '1. This Regulation lays down rules relating to the protection of natural persons with regard to the processing of personal data and rules relating to the free movement of personal data. 2. This Regulation protects fundamental rights and freedoms of natural persons and, in particular, their right to the protection of personal data. 3. The free movement of personal data within

[247] Robertson, A.H. (1975). Collected edition of the 'travaux préparatoires' of the European Convention on Human Rights = Recueil des travaux préparatoires de la Convention Européenne des Droits de l'Homme; Council of Europe, vol. 2 Consultative Assembly, second session of the Committee of Ministers, Standing Committee of the Assembly, 10 August-18 November 1949, 270 (The Hague: Nijhoff).
[248] Blok, P. H. (2002). Het recht op privacy: een onderzoek naar de betekenis van het begrip 'privacy' in het Nederlandse en Amerikaanse recht. Den Haag.
[249] Van Dijk, P., Hoof, G. J., & Van Hoof, G. J. (1998). Theory and practice of the European Convention on Human Rights. Martinus Nijhoff Publishers.
[250] Evans, C. (2001). Freedom of religion under the European Convention on Human Rights (Vol. 1). Oxford University Press.
[251] Protocol No. 9 to the Convention for the Protection of Human Rights and Fundamental Freedoms Rome, 6.XI.1990. This Protocol has been repealed as from the date of entry into force of Protocol No. 11 (ETS No. 155) on 1 November 1998. Protocol No. 11 to the Convention for the Protection of Human Rights and Fundamental Freedoms, restructuring the control machinery established thereby. Strasbourg, 11.V.1994. Since its entry into force on 1 November 1998, this Protocol forms an integral part of the Convention (ETS No. 5).
[252] It does allow citizens that all have been harmed by a specific governmental practice to bundle their complaints.
[253] ECtHR, *Church of Scientology of Paris v. France*, application no. 19509/92, 09 January 1995.
[254] See e.g. ECtHR, *Lawlor v. The United Kingdom*, application no. 12763/87, 14 July 1988. ECtHR, *Tauira and others v. France,* application no. 28204/95, 04 December 1995. ECtHR, *Asselbourg and 78 others and Greenpeace Association-Luxembourg v. Luxembourg,* application no. 29121/95, 29 June 1999.

the Union shall be neither restricted nor prohibited for reasons connected with the protection of natural persons with regard to the processing of personal data.'

One of the explicit goals of the 1995 EU data protection framework was to remove obstacles to the transfer of personal data within the European Union by laying down one common level of data protection. One of the problems that existed before the EU data protection framework was put in place was that each country had different data protection standards embedded in its national laws. This hampered the use and transfer of personal data, as a company could only transfer personal data from Germany to Italy if it ensured that it respected both the Italian and German data protection regimes. Companies operating in all EU countries had to comply with a different legal regime for each data processing operation, which created many barriers for internationally operating companies, not least because different national laws sometimes imposed conflicting requirements. Adopting a single EU-wide data protection framework eliminated restrictions on trade and data transfer while ensuring a high level of data protection.

In addition, it is important that the rules in the GDPR seldom prohibit specific data processing operations. In most cases, they lay down procedural safeguards and principles ensuring accurate and secure data processing operations. That is why some see the data protection regime not so much from a protective angle, but as laying down rules ensuring good and fair processing operations, thus stimulating data innovation and use, rather than limiting it.

Finally, the GDPR contains many explicit exceptions for specific processing operations. Most important for this study are the exceptions in relation to the freedom of speech, archiving, statistical research, open government and the re-use of public sector information. It is important to signal the push within the EU for open data and the re-use of public sector information. Traditionally, Western society has been based on the belief in an open and transparent government. The idea is that critical citizens and journalists should be able to assess decision-making processes both in light of active citizenship and in order to expose potential flaws. Open government is therefore believed to be part and parcel of a vital democracy.

The EU has made a choice to go beyond promoting openness and transparency vis-à-vis governmental practices; it has also stimulated the re-use of government information. The underlying belief is that the government is sitting on 'a mountain of data', while its economic potential is not being fully exploited. If the data were released for the commercial re-use, it is estimated that tens of billions in economic potential would be released in the European Union alone.[255] The EU, therefore, adopted a PSI re-use directive in 2003, which,[256] following amendments in 2013[257] and 2019,[258] has become even more forceful in encouraging member states to actively release PSI for re-use by commercial parties. Still, the Open Data Directive makes clear that it does not affect the GDPR.

Interestingly, just like the issues described in section 6.2, there is ambiguity within the EU on how to deal with the conflicts between the various regimes and the rationales underlying them. In section 6.2, it was shown that, in general, the EU regulator is set on issuing a regulation that is based on clear and separate data categories, while the courts have adopted more contextual and fluid approaches. Advisory bodies such as the Article 29 Working Party and the European Data Protection Board also propagate a flexible approach and have stretched and broadened the scope of, inter alia, the concept of personal data over time. In addition, courts have set clear limits when regulators use data distinctions to adopt lower levels of protection. The CJEU has declared null and void the Data Retention Directive, that required

[255] Dos Santos, C., et al. (2013). 'LAPSI Policy Recommendation n. 4: Privacy and Personal Data Protection', Lapsi Recommendation.
[256] **Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information.**
[257] Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information.
[258] Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information.

Member States to store metadata for long periods of time; the ECtHR has adopted a complex web of rules for intelligence agencies that gather and analyse large amounts of metadata.

A similar approach can be witnessed with respect to the move towards open data and the re-use of public sector information. While this is highly stimulated by the EU regulator, the courts are more hesitant. For example, with respect to a regulatory regime adopted in Latvia, the CJEU questioned whether, in order to protect or improve road safety, granting access to data about traffic violations was the least intrusive. It found that the regime allowed third parties to access the information even if those third parties had other purposes than those related to increasing road safety. This is not allowed because of the purpose limitation principle, the Court pointed out. Thus, when open data are the re-use of public sector information initiatives entail the processing of personal data, they must conform in full to the privacy and data protection regime. This means that there should be limits to, inter alia, the re-use of information for other purposes than for which they were gathered, which is generally impossible when data are made available online and, in any case, goes against the very idea of promoting re-use for commercial purposes.

6.4 Bottlenecks and dangers of under- and over-regulation

The difficulty of assessing the existence of regulatory gaps and the desirability of regulatory alternatives is that the discussion on the regulatory goal(s) of the privacy and data protection framework should be settled first, while that remains a matter of debate. In addition, there is no preferred regulatory approach: a categorical, a contextual or a hybrid one. Each has its own set of advantages and disadvantages. Consequently, it is a matter of perspective whether regulatory gaps exist and, if so, how they should be remedied. In addition, choosing between different regulatory options entails a choice of where to put the regulatory prerogative. The more clarity is provided in the legal regime, the more the prerogative is put with the legislative branch. The more a contextual approach is taken, the more the correct interpretation of the rules per context has to be given by the judicial and/or the executive branch. The former has the advantage of democratic legitimacy, and the latter has the advantage of practical validity. The former has the advantage of providing legal certainty by giving one approach to all situations alike, and the latter has the advantage of being able to provide regulatory granularity.

To give an example, perhaps the essential question this study raises is whether the notion of ‘personal data’ should be retained and whether non-personal data should be provided with a form of protection, for example, under a GDPR light regime. That question is dependent on what the regulatory rationale of the data protection framework is believed to be. If it is providing protection to natural persons’ individual interests, then there is no direct need to also regulate the processing of aggregated or anonymous data. In order to tackle potential harms that arise from data policies and practices based on group profiles, it might be left to the courts to interpret the regulatory regime so as to cover those harms either on the basis of the GDPR or under Article 8 ECHR. If the regulatory objective is that of curtailing data power by public and private sector organisations, then it both makes sense to also set limits on and requirements for the processing of non-personal data, and it would be no problem to expand the data protection regime to also cover the processing of non-personal data.

Both choices, in addition, beg the question of regulatory specificity. The regulator has so far maintained a strict regulatory distinction between non-personal and personal data, yet in practice, this distinction is difficult to draw. Courts have consequently expanded the definition of personal data to cover data that is more and more peripheral to the natural person, while data controllers are asking for more cues on how to make that distinction. The danger of leaving the current approach intact is that responsible data organisations will err on the safe side, while others will explicitly seek the boundaries of the law. In addition, the less regulatory clarity is given, the more difficult it will be to enforce the rules because every data processing operation might require its own assessment of legality and legitimacy. Consequently, when the choice is made to keep the current regulatory regime intact, the question is still

whether more regulation should be provided to data controllers on how to draw distinctions between data categories.

In addition, when the choice is made to cover non-personal data. Two different approaches can be taken: a categorical and a contextual one. Either the regulatory regime maintains a difference between non-personal and personal data but attaches a different regulatory regime to non-personal data, or it does away with this differentiation and potentially other data distinctions and puts the type of rules and the regulatory burden on data controllers fully dependent on the case by case assessment of the risks involved (when this approach is taken, obviously it matters whether the risks are related to individual, group and/or societal interests).

Then, for the question of overregulation, it matters to what extent stimulating data processing operations is set on the same foot as the protective rationale of the data protection regime and how the goal to promote open data environments and the re-use of public sector information is evaluated. Should the latter rationale be seen as an equally important rationale as the protective rationale, or can this rationale only be furthered within the boundaries set by the protective rationale? If the latter is the case, overregulation is not an important risk, while avoiding underregulation is the main objective. If, however, the rationales are set on the same foot, furthering one has an impact on the other, laying strict rules for sharing open data may serve the protective rationale but may undermine the goal of stimulating data processing operations.

In addition, when promoting data processing is seen as an important rationale that sits on the same level as the protective rationale, the question is still what type of regulation would be most effective. Although an open and contextual framework seems to leave the most room for data innovation at first sight, data controllers often call explicitly for more regulatory clarity and certainty because they fear backlashes and investments that don't pay off. Although a strict regulatory regime might seem stifling, at first sight, it might, in fact, lead to data operations that have broad support among the population, regulators and other players and so be desirable in light of stable business growth. This matter of regulatory effectiveness is not something this study has assessed, but it should play an important role in the evaluation of the desirability of introducing alternatives to the current data protection regime.

A similar point should be noted with respect to the protective rationale. Experts have stressed that the approach taken by the GDPR, under which processing sensitive data is in principle prohibited, is increasingly missing the goal it sets out to achieve, namely to protect individuals against harm. In order to prevent discriminatory practices in AI systems, it may be necessary to process sensitive personal data. Consequently, rather than prohibiting processing sensitive data, the data protection regime should perhaps mandate processing such data in the AI context. Others have stressed that in order to serve the protective rationale of privacy and data protection regimes in the AI context, it may be necessary to focus not on data minimisation but on data minimisation, on requiring a minimum level of data to be gathered, analysed and stored rather than a minimal level.[259] Thus, it is a matter of debate whether the protective rationale is best served by laying down limitations on data processes.

A final point that should be mentioned is the fact that regulation needs not only come from the legislative side.[260] Social norms may develop, and societal practices may emerge, changing what is perceived to be normal, acceptable or desirable in terms of data processing. Even if the argument would be true that societal norms have so far only evolved toward being more acceptant of immersive data technologies, the question is whether this is problematic. In addition, privacy-preserving technologies are being developed that gain ground slowly but surely. An example, but certainly not the only one, is the increased popularity of the search engine Duckduckgo. Privacy-preserving competitors could in time

[259] Van der Sloot, B. (2013). From data minimization to data minimummization. In *Discrimination and Privacy in the Information Society*. Springer, Berlin, Heidelberg, 273-287.
[260] Lessig, L. (2009). Code: And other laws of cyberspace.

for alternatives to the Big Five. Though it is difficult to imagine their data power being broken now, the market is still relatively young and changes significantly every year.

6.5 Regulatory alternatives

In academic literature and policy reports, a high number of regulatory alternatives have been suggested to alter the current regulatory framework in order to close the existing gaps between the legal and the technological realm. This section will provide an overview of the most important suggestions for the purposes of this study. This will be divided along the lines of the four data categories that were central to this study: anonymous data (section 6.5.1), aggregate data (section 6.5.2), pseudonymous data (section 6.5.3) and sensitive personal data (section 6.5.4).

6.5.1 Anonymous data

Seeing the tensions summarised in section 2.4, scholars have suggested regulatory alternatives roughly along three lines: keeping the current framework with amendments, including under the framework the (partial) regulation of non-personal data, and looking for new ways to delineate concepts.[261]

A first option would be to essentially do away with the distinction between anonymous and non-anonymous data. If it is accepted that tools and techniques for (re)identification are or will be so advanced that anonymization is no longer possible or feasible, given the costs and effort involved, the choice to place anonymous or non-personal data outside the scope of data protection law would be redundant. That is why it has been suggested to apply a GDPR-light regime to non-personal data.[262]

A second option would be to use a narrower concept than personal data to distinguish more clearly between personal and anonymous information, as information would then be considered anonymous for a longer time. For example, in the USA, the term personally identifiable information is used, which can be argued to be narrower in scope than the concept of personal data under EU law. Similar would be to use a concept of 'depersonalization' stricto sensu versus anonymisation, in a sense that depersonalization could then refer to, for example, information that is stripped of identifiers but not completely anonymous. The latter is a bit similar to applying measures such as pseudonymization.

A third option is to create different levels when it comes to identifiability. Schwartz and Solove, for example propose to keep personal information as a threshold of protection but using a sharper definition. They propose to link the concept to the risk of identification, ranging from '0' to 'identified'. For the different levels of identifiability in the information, there would be different requirements.[263] This links well to the contextual approach of the data protection regime. Finck and Pallas essentially describe anonymization as risk management.[264] Guidance would be needed on the elements that define the different levels, how to make dynamic levels that grow with the technological possibilities, and which legal safeguards to require for which levels.

A fourth option is to take a context-dependent approach to anonymization. For example, after the Breyer case, Stalla-Bourdillon and Knight proposed a dynamic and fluid approach to anonymisation: 'data

[261] See for example: Rezlauf, I. (2020). EU Framework for Handling Big Datasets Mixed of Personal and Non-personal Data. Computer Law Review International, 21(1), 7-13; Spindler, G., & Schmechel, P. (2016). Personal data and encryption in the European general data protection regulation. J. Intell. Prop. Info. Tech. & Elec. Com. L., 7, 163. Purtova, N. (2018). The law of everything. Broad concept of personal data and future of EU data protection law. Law, Innovation and Technology, 10(1), 40-81.

[262] van der Sloot, B. (2016). The individual in the big data era: moving towards an agent-based privacy paradigm. Exploring the boundaries of Big Data, 177. Van der Sloot, B. (2017). Privacy as virtue: Moving beyond the individual in the age of big data (Vol. 11). Intersentia. van der Sloot, B. (2020). Regulating non-personal data in the age of Big Data. In Health data privacy under the GDPR: Big Data challenges and regulatory responses (pp. 85-105). Routledge.

[263] Schwartz, P. M., & Solove, D. J. (2011). The PII problem: Privacy and a new concept of personally identifiable information. NYUL rev., 86, 1814.

[264] Finck, M., & Pallas, F. (2020). They who must not be identified—distinguishing personal from non-personal data under the GDPR. International Data Privacy Law.

should always remain a fluid line-anonymized data can always become personal data again depending upon the evolution of the data environment. Said otherwise, a dynamic approach to anonymized data is warranted. In this sense, the opposition between anonymized data and personal data is less radical than commonly described. Such a dynamic and thereby contextual approach to anonymized data is compatible with the new data protection regime to be found in the GDPR.'[265] A dynamic and fluid approach requires case-by-case assessments and constant reflection.

A fifth option is a functional approach to anonymization, which focuses on the environment in which the data is situated. Elliot et al. propose this approach, explaining that 'if the failure of anonymisation is down to uncertainty about the auxiliary information, it follows that one cannot tell from the data alone whether a dataset is anonymous, for the obvious reason that the data alone says nothing about the auxiliary data.'[266] According to them, a contextual approach is important but should not revolve around technologies. Functional anonymization is defined as: 'whether data is anonymous or not (and therefore personal or not) is a function of the relationship between that data and its environment.' The environment is determined by four few elements: other data, data users, governance processes, and infrastructure. Related to this is the idea of regulating the process or the outcome rather than having a definition of anonymous data or distinguishing the concept of anonymous data itself.

6.5.2 Aggregate data

Seeing both the tension between the legal and the technical domain and between the legal domain of privacy and data protection law, and that of open data and re-use of public sector information, as discussed in section 3.4, several regulatory alternatives have been suggested in the literature.

First, a study on SDC and the GDPR concluded that the growth of data increases the threat of personal data disclosures. On the one hand, an intrinsic risk is the growth in size of a dataset, which makes it difficult to detect and deal with data disclosure risks that are hidden in the datasets. On the other hand, an extrinsic risk is the growth in number of other datasets available to other parties makes it difficult to assess and deal with the data disclosure risks that may arise when combining datasets. These risks make it difficult for data controllers to share their data with specific groups, individuals, or the public. Thus, an option could be to have the data protection regime prevail over or provide the main framework for using, sharing, and making public statistical and aggregated data.[267] Essentially, this is what the CJEU proposed in *Latvijas Republikas Saeima*.[268]

Second, authors have suggested an extensive framework to reconcile the need for open data and processing statistical data on the one hand and the need for privacy and data protection on the other. A case-by-case assessment of the rules and regulations would be necessary under this framework and can be guided by a circumstance catalogue containing questions that reflect on possible risks of the disclosure. In that context, they distinguish between four data categories according to risk levels: raw personal data, pseudonymous data, anonymized data, and non-personal data. Those four categories are defined as following: 'With raw personal, no attempt has been made to make identification harder. Pseudonymous data are data for which the individual's name is changed to another unique identifier. Anonymized data are ex-personal data; people cannot be re-identified in the dataset. Non-personal data, such as data about weather conditions or public transport times, never contain personal data.'[269] For the raw personal data, they propose to not release those as open data, non-personal data generally to be

[265] Stalla-Bourdillon, S., & Knight, A. (2016). Anonymous Data v. Personal Data False Debate: An EU perspective on Anonymization, Pseudonymization and Personal Data. Wisconsin International Law Journal, 34(2), 284-322.
[266] Elliot, M., O'hara, K., Raab, C., O'Keefe, C. M., Mackey, E., Dibben, C., ... & McCullagh, K. (2018). Functional anonymisation: Personal data and the data environment. Computer Law & Security Review, 34(2), 204-221.
[267] Van der Sloot, B. (2011). Public sector information & data protection: A plea for personal privacy settings for the re-use of PSI. Informatica e Diritto, Fascicolo, 1-2.
[268] See for a different approach: Folmer, E., & Paapst, M. (2019). Linked open data & AVG: niks aan de hand, of toch wel?. *Geo-info*, (3), 12-14.
[269] Borgesius, G. & Van Eechoud, M. (2015). 'Open Data, Privacy, and Fair Information Principles Towards a Balancing Framework', Berkeley Technology Law Journal, Vol. 30(3) 2073-2131.

released without constrictions, and for pseudonymous and anonymous data to set restrictions for release. As for anonymized data, there is still a risk of re-identification. Those datashould only be disclosed with access and re-use restrictions. For restrictions, ideas would be licensing for the data access, on-site access, etc.

Third, a possible technical solution for the use of statistical data while safeguarding the interests of those represented in the data, more specifically to prevent their personal information from being disclosed, can be found in a technique such as statistical disclosure control. It is argued that SDC technologies can be used for achieving data minimization, purpose limitation, and proportionality under the GDPR. A threshold could be introduced to mark the boundary of data anonymity under the GDPR. The threshold would depend on the context and time, for example, taking into account available technologies, other available data sources, and the motivations for and costs of reidentification. That threshold thus fluctuates and is contextual: currently, anonymous data may become non-anonymous personal data as the anonymity threshold rises or the threshold level may lower when the current background knowledge does not longer exist.[270] This would be interesting to align with the thresholds that NSIs use to mark for different types of files or datasets when they can be disclosed and are thus not disclosing personal information.

Fourth, a more radical alternative could be to find ways to base privacy and data protection regulation on other concepts than identifiability. For example, some authors have suggested leaving the (sole) focus on individual privacy and linkability aside and instead or in addition to focusing on groups, categories, and data collectives. A term that has been coined is group privacy: 'The search for group privacy can be explained in part by the fact that with big data analyses, the particular and the individual is no longer central. In these types of processes, data is no longer gathered about one specific individual or a small group of people, but rather about large and undefined groups. Data is analysed on the basis of patterns and group profiles; the results are often used for general policies and applied on a large scale. The fact that the individual is no longer central, but incidental to these types of processes, challenges the very foundations of most currently existing legal, ethical and social practices and theories. [] Although this focus on personal identifying information is still useful for more traditional data processing activities, it is suggested by many that in the big data era, it should be supplemented by a focus on identifying information about categories or groups.'[271]

6.5.3 Pseudonymous data

Given the tensions between the legal and the technological realm as discussed in section 4.4, several regulatory alternatives have been proposed.

First, ENISA proposes that controllers and processors should engage in data pseudonymisation based on a risk assessment taking account of the overall context and characteristics of the personal data processing, including methods for data subjects to pseudonymise personal data on their side.[272] While it is an interesting idea to involve the data subjects, at the same time, it raises the question of how data subjects would be able exactly to pseudonymise data on their side and if they have enough knowledge of the process and of potential harms if they do not do so.

Second, ENISA also recommends that, for example, the Data Protection Authorities and the European Data Protection Board should promote risk-based data pseudonymisation and provide guidance and

[270] Bargh, M. S., Meijer, R. F., & Vink, M. (2018). On statistical disclosure control technologies: For enabling personal data protection in open data settings. WODC Cahiers,20.
[271] Taylor, L., Floridi, L., van der Sloot, B. eds. (2017) Group Privacy: new challenges of data technologies. Dordrecht: Springer.
[272] ENISA, Data pseudonymisation: advanced techniques &use cases. Technical analysis of cybersecurity measures in data protection and privacy. January 2021.

examples for controllers and processors. Which technique should be used depends on the context and which technique is most suitable and sufficient.

Third, ENISA proposes to align pseudonymisation with the concept of data custodianship. A data custodian could function as a Pseudonymisation Entity (PE), the entity responsible for processing identifiers into pseudonyms using the pseudonymisation function (which can be a controller or processor), that can allow data access under specific conditions to researchers or companies in an interconnected data ecosystem and on the other hand for shielding data against unwanted or unlawful access. The data custodian can fulfil different roles: it can fulfil the role of assigning pseudonyms by applying the pseudonymisation function to the identifying data, or store pseudonymised data being provided by the data controller and facilitating access after the authorised parties have proven legitimacy, or provide synthetic data.[273]

Fourth, while there is the criticism of linkability easily being restored within pseudonymous data, there are also arguments to be made that such weak pseudonymisation would not be sufficient to meet the concept under the GDPR: 'Pseudonymization' commonly refers to a de-identification method that removes or replaces direct identifiers (for example, names, phone numbers, government-issued ID numbers, etc.) from a data set, but may leave in place data that could indirectly identify a person (often referred to as quasi-identifiers or indirect identifiers). Applying such a method, and nothing else, might be called 'simple pseudonymization.' Frequently, security and privacy controls designed to prevent the unauthorized reidentification of data are applied on top of simple pseudonymization to create 'strong pseudonymization.'[274] In this light, it can be argued that the GDPR is too open and vague in its approach to pseudonymous data. Not only does it not stipulate requirements regarding the technical process be used, but some underline it also does not distinguish between different situations and actors. On these points, the GDPR could be revised to ensure more clarity.

Fifth, one interpretation is that the GDPR affords protection against internal reidentification: the technical and organizational measures required refer to the *'additional information*' which must be '*subject*' to these measures, thus, the additional identifiable information held separately from the pseudonymised data must be protected. In this interpretation, the only risk of identification mitigated by pseudonymisation is the risk of identification through the original data held either by the controller or by a third party.[275] This would exclude re-identification by other means. Some conclude that there can be a controller of the pseudonymized data who is in possession of a separately kept re-identification mechanism, or there can be a third party accessing the data, while recital 29 does not distinguish between these actors while there is a greater risk of re-identification with the controller.[276] They suggest that the GDPR should distinguish between the transfer to and use of the data by third parties (controllers who have no access to the decryption algorithm) and controllers who, through separate means, are in possession of the re-identification means by awarding privileges to the former and imposing punishments for re-identification into the conditions for privileged use of pseudonymized data.[277]

Sixth, it is argued that 'pseudonymization can be used both to reduce the risks of reidentification and help data controllers and processors to respect their personal data protection obligations by keeping control over their activities. On the one hand, pseudonymization ensures the capability to reconstruct the processes of identity masking, by allowing re-identification. On the other hand the accountability of the data controller and data processor is guaranteed, thanks to the fact that there will always be a person

[273] See further: Van der Sloot, B., & Keymolen, E. (2022). Can we trust trust-based data governance models? In search of a regulatory model for implementing data trusts. Data and Policy.
[274] Hintze, M., & El Emam, K. (2018). Comparing the benefits of pseudonymisation and anonymisation under the GDPR. Journal of Data Protection & Privacy, 2(2), 145-158.
[275] Mourby, M., et al. (2018). Are 'pseudonymised'data always personal data? Implications of the GDPR for administrative data research in the UK. Computer Law & Security Review, 34(2), 222-233.
[276] Kotschy, W. (2016). The new General Data Protection Regulation-Is there sufficient pay-off for taking the trouble to anonymize or pseudonymize data.
[277] Koot, M. R. (2012). Measuring and predicting anonymity.

who can re-identify subjects included in a cluster, acting as a "data keeper".'[278] This argument is an argument in favour of maintaining the category of pseudonymous personal data as an in-between category between non-personal and personal data. In general, the attitude towards pseudonymous data as a useful category seems to be much more optimistic compared to distinguish between anonymous and non-anonymous data, as pseudonymization offers incentives for protecting data in terms of exceptions from legal obligations while still falling within the scope of the GDPR thus not creating a major risk of harm to those represented data subjects in the data.[279]

Finally, some have propagated further emphasis on MPC, which would have as benefit over techniques such as differential privacy that there is no trade-off between privacy and accuracy. Compared to the techniques of using a third party, MPC has the benefit that it removes the need for such a party and works in a decentralized way. Nonetheless, the use of MPC is also not without its own challenges. For example, if the adversary could access enough executions of the protocol, it could identify individuals. If the adversary has access to few executions, it would not be statistically probable for him to crack the protocol; however, if the number of executions grows, that risk increases. Perfect security would protect the system against adversaries with unlimited computing resources and time but it is not always achievable. MPC can be used as a pseudonymization or anonymization technique for the processing and sharing of personal data. It is conceived as a privacy by design and default tool and an appropriate technical and organizational measure. Finally, the system is not vulnerable to computationally powerful adversaries in many cases, and it is less computationally expensive and complex than other techniques, such as fully homomorphic encryption. The question remains as to MPC could be systematically considered as an anonymization or pseudonymization technique. Here, a case-by-case analysis would render the correct result. It is clear, though, that MPC offers more benefits than most other techniques.

6.5.4 Sensitive personal data

Given the tensions between the legal and the technological realm as discussed in section 5.4, several regulatory alternatives have been proposed in the literature:

First, Hildebrandt and other scholars propose that the regime for special categories of data is no longer adequate in the era of big data analytics. This is so because the same data may be sensitive in one context but not in another (particularly where data are combined), making it more unclear whether specific categories of data are sensitive as the use of these data may or may not be sensitive depending on each context.[280] Van der Sloot and Van Schendel have put the distinction between sensitive and non-sensitive data under the legal regime in broader perspective and have argued that big data processes would increasingly challenge legal regimes that work with static categories of data: 'While the current legal system is focused on relatively static stages of data, and linked to them specific forms of protection (e.g. for personal data, sensitive data, private data, statistical data, anonymous data, non-identifying information, metadata, etc.), in reality, data go through a circular process: data is linked, aggregated and anonymized and then again de-anonymized, enriched with other data and profiles, so that it becomes personally identifying information again, and potentially even sensitive data, and is then once again pseudonymised, used for statistical analysis and group profiles, etc.'[281] Zarsky has seconded that

[278] Bolognini, L., & Bistolfi, C. (2017). Pseudonymization and impacts of Big (personal/anonymous) Data processing in the transition from the Directive 95/46/EC to the new EU General Data Protection Regulation. Computer law & security review, 33(2), 171-181.
[279] On the advantages of pseudonymization see for example: Schwartmann, R., & Weiß, S. (2017). White Paper on Pseudonymization Drafted by the Data Protection Focus Group for the Safety, Protection, and Trust Platform for Society and Businesses in Connection with the 2017 Digital Summit. Digit. Summit, 2017, 44; Hintze, M., & El Emam, K. (2018). Comparing the benefits of pseudonymisation and anonymisation under the GDPR. Journal of Data Protection & Privacy, 2(2), 145-158.
[280] Hildebrandt, M. (2009). Who is profiling who? Invisible visibility. In Reinventing data protection? (pp. 239-252). Springer, Dordrecht; Politou, E., Alepis, E., & Patsakis, C. (2019). Profiling tax and financial behaviour with big data under the GDPR. Computer law & security review, 35(3), 306-329; Cockfield AJ. Big data and tax haven Secrecy. Fla Tax Rev 2015;18:483.
[281] Van der Sloot, B., & van Schendel, S. (2016). Ten questions for future regulation of big data: A comparative and empirical legal study. J. Intell. Prop. Info. Tech. & Elec. Com. L., 7, 110.

finding: 'If nearly all forms of data categories and data sets can produce special data, why even bother with this distinction, which is rendered almost artificial?'[282]

Second, suggestions have been made to broaden the list of sensitive data and include in it, inter alia, financial data, as was also discussed when drafting the GDPR but was ultimately rejected. Privacy scholars have repeatedly highlighted that tax and financial data are considered to be among the most sensitive forms of personal information, as they may reveal, among others, information about income, spending and savings, employment status, person's health, marital status, lifestyle, hobbies, personal belongings, and disability status. As financial data provides insights into these types of characteristics, it is particularly useful for creating profiles, including religious and political beliefs, political alliances, and personal behaviour.[283] Article 29 Working Party stated that personal data linked to taxes might be deemed as sensitive data and, therefore, care should be taken to afford it higher standards of data protection.[284]

Third, Quinn and Malgieri propose an interpretative solution: a hybrid approach where a purpose-based definition acquires a bigger role in deciding whether data is sensitive, combined with a context-based 'backstop' based on reasonable foreseeability. 'A purpose-based interpretation of sensitive data, with a relevant context-based backstop. In other words, personal data should be considered sensitive IF the intention of the data controller is to process or discover sensitive information OR if it is reasonably foreseeable that, in a given context, the data in question can be used to reveal or to infer sensitive aspects of data subjects. This formulation would have the advantage of not only seeing data as sensitive where there was an intention of processing sensitive data or a real risk of doing so but would simultaneously avoid the label of sensitive data being applied where there was no intention to process sensitive data and where there was no reasonably foreseeable prospect that this could be the case. The authors of this Article would argue that it is only through such a formulation that a balance can be struck where the concept of sensitive data remains viable, and a real level of protection is offered to data subjects who may be in a vulnerable position and at risk from discrimination and associated phenomena in line with their fundamental rights.'[285]

Fourth, while most concerns are about whether the GDPR is strict enough on special categories of data or not, there are also arguments to consider from the opposite perspective. There is an ongoing discussion on to what extent it is possible to process sensitive personal data in order to prevent discrimination, for example, in AI systems. Zliobaite and Custers propose that using sensitive personal data may be necessary for avoiding discrimination, especially when it comes to data-driven decision-making.[286] Thus in order to further one of the underlying rationales of the category of special data, namely to prevent discriminatory practices, it may be necessary to process sensitive personal data instead of the other way around.

6.6 Scenario's

Given the main findings of this study as outlined in section 6.2, put in light of the various choices discussed in sections 6.3 and 6.4, and seeing the high number of specific regulatory alternatives as mapped in section 6.5, it is not possible to give a set of recommendations. What is an additionally complicating element to this task is that it is unsure how the data landscape will evolve over time. At

[282] Zarsky, T. Z. (2017). Incompatible: the gdpr in the age of big data. Seton Hall Law Review, 47(4), 995-1020.
[283] Politou, E., Alepis, E., & Patsakis, C. (2019). Profiling tax and financial behaviour with big data under the GDPR. Computer law & security review, 35(3), 306-329; Cockfield AJ. Big data and tax haven Secrecy. Fla Tax Rev 2015;18:483; Sharman, J. C. (2009). Privacy as roguery: Personal financial information in an age of transparency. Public Administration, 87(4), 717-731.
[284] Politou, E., Alepis, E., & Patsakis, C. (2019). Profiling tax and financial behaviour with big data under the GDPR. Computer law & security review, 35(3), 306-329.
[285] Quinn, P., & Malgieri, G. (2021). The Difficulty of Defining Sensitive Data—The Concept of Sensitive Data in the EU Data Protection Framework. German Law Journal, 22(8), 1583-1612.
[286] Zliobaite I. & Custers B. (2016), Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models, Artificial Intelligence and Law (24): 183-201.

least two aspects are relevant in this respect: the availability of data, especially through open access to online data sources, and technological developments. The exact way in which both aspects will develop over time is not clear and difficult to predict.

If more and more data are available, the chances that it will be possible to re-identify a person in an anonymised dataset or that two aggregated datasets can be merged to derive personal data increases. The same applies to the possibility to distil content data from metadata and sensitive data from non-sensitive data. The rise of new data technologies may additionally mean that the underlying rationales of the various data categories are challenged because the processing of non-personal data may have just as a big an impact or an even bigger one than when an organisation processes personal data because bulk interception of metadata may be just as or even more impactful than collecting content communications data, etc. What is more, data technologies may emerge that are highly invasive, yet are not covered by the data protection framework in general or the category of sensitive data specifically. The list of sensitive data has already been expanded in the GDPR, inter alia, to include biometric data, but given the pace at which technologies develop, the list may need to be updated more regulatory.

Both developments mean that the very idea of working with distinct data categories is increasingly difficult to uphold because data are no longer stable in nature. A dataset containing ordinary personal data can be linked to and enriched with another dataset so as to derive sensitive data; the data can then be aggregated or stripped of identifiers and become non-personal, aggregated or anonymous data; subsequently, the data can be deanonymized or integrated into another dataset in order to create personal data. All this can happen in a split second. This is already the case; if both the availability of data increases and the capacity of data technologies expands, it might become simply impractical to work with fixed statuses of data.

Given these factors, it is possible to distinguish between five hypothetical scenarios and regulatory options that are connected to those scenarios. These are ideal type scenarios.

Scenario 1: leaving the data protection framework as is

In the first scenario, the data protection framework is regarded as forming a perfect equilibrium between its protective rationale and the rationale promoting data processing operations, between opting for a categorical and a contextual regulatory approach and between leaving the regulatory prerogative to the legislator and allowing judicial and executive authorities to refine concepts and rules in practice, with an eye to specific contexts and situations. Though the technological practice may be said to diverge from the regulatory regime and may very well do more so over the years, this does not mean that the rules should change. Rather, more should be invested in ensuring that practice is kept in conformity with the rules.

To the extent that processing non-personal data has an important impact, such is already covered by the GDPR when decisions are taken in which a person is singled out or significantly affected, or by Article 8 ECHR, when policies affect the very broad notion of private life as interpreted by the European Court of Human Rights. In addition, the jurisprudence of the ECtHR has shown that it is both willing to develop a regime for metadata collection when necessary and to allow claims in which no personal harm was endured by the claimant, instead focussing on the societal effects of large-scale data processing. Data protection law does not need to solve all problems of the data-driven environment.

Thus, the current regulatory regime is ready for the 21st century. Although it is true that neither one of the sources consulted for this study confirmed this scenario, it may be argued that there is no consistency in the regulatory alternatives suggested. For example, while some argue for more contextuality, others argue for more clearly defined categories. Instead of favouring one alternative over the other, the current regulatory regime in scenario 1 keeps both groups equally (dis)satisfied.

Scenario 2: keeping the data protection framework and investing in more precise rules

In scenario two, the main outlines and contours of the current regulatory regime are deemed fit for the 21st century, but the main regulatory challenge is seen in the need for further clarity on the definitions of the different data categories, the boundaries between different categories and the regulation of those types of data. Under this scenario, various regulatory alternatives are possible.

First, there could be a push for more guidelines and best practices on the correct interpretation of the data protection framework. It is clear from this study that many experts hope the European Data Protection Board would issue several detailed opinions, not only on the boundary between personal and non-personal data or between sensitive and non-sensitive personal data, but also on the various techniques that may or should be used in light of anonymisation and pseudonymisation.

Second, more clarity could be provided on how to determine whether a dataset is, in fact, anonymous. For example, a rule could be introduced laying the burden of proof on data controllers and/or mandating that the data controller have external hackers perform attacker scenarios on their database and hire experts to check to assess that data cannot be reasonably re-identified. Similarly, rules can be set on how to ensure that data is, in fact, pseudonymous. A suggestion that came up during this study is to appoint a special data custodian in an organisation that has control over the pseudonymisation process and has the key to the data.

Third, to provide more clarity on the distinction between non-personal and personal data, the contextual elements in the definition of personal data (identifiable) and in the description of anonymisation (To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments) could be removed. This would decontextualise the question of whether personal data are processed and whether the data protection framework applies.

Fourth, the category of pseudonymous data can be deleted. This category is critiqued both for its vagueness and because it privileges one privacy-preserving technique over others, for which no clear explanation exists.

Fifth, it may be considered to the extent the list of sensitive personal data. Potential additional categories that were identified in this study include financial and socio-economic data, data about children, locational data and even metadata.

Finally, a small but radical change would be to the extent the notion of identifiability so that it not only covers the identifiability of natural persons. Already, the e-Privacy regime provides protection to legal persons as well as natural persons. In addition, it could be considered to have the GDPR also cover the identifiability of groups. Listing these three categories would put the GDPR in line with the European Convention on Human Rights, which attributes rights to natural persons, non-governmental organisations and groups and ensure that a big regulatory gap that is identified by many in this study is closed, namely the fact that processing aggregated data and developing group profiles falls outside of the protective scope of the current regime. This regulatory alternative would seem obvious when the protective rationale of data protection is connected to the protection of societal interests and putting due process requirements on organisations with data power, while it may be considered to lead to overregulation when the protective rationale is linked to the interests of natural persons.

Scenario 3: keeping the data protection framework and investing in more contextuality

Scenario 3 is like scenario 2, only the main regulatory challenge is regarded as the lack of contextuality and adaptability of the current regulatory regime. Again, several regulatory alternatives have emerged during this study.

First, it can be considered to introduce in the data protection framework, specifically in the list of principles in article 5, a principle of contextuality. This principle would require the controller to consider each principle, obligation and requirement it has to adhere to under the data protection framework in light of the context in which the data process takes place and has an effect.

Second, it could be considered to reformulate the list of sensitive data in the way it was originally formulated, namely as examples rather than an exhaustive list. Alternatively, it could be considered to include a residual category, similar to Article 14 ECHR.

Third, the category of pseudonymous data could be given a more prominent role, providing a clear grey category in between the black and white categories of personal and non-personal data. While there are some limitations to the data protection regime that apply to pseudonymous data, these are still relatively minor exceptions, which provide insufficient incentives to invest heavily in strong pseudonymisation. This could be changed by giving pseudonymous data a stronger position in the data protection framework.

Finally, a sectoral approach could be considered. Europeans used to mock Americans for their sector-specific approach to data protection; they had informational privacy standards for specific domains, such as laws for the protection of online privacy of children, laws concerning privacy protection in the health care sector, laws regarding data processing in light of credit reporting, etc. Europe, instead, had an omnibus law, that applied to all data processing activities irrespectively. Thus, there were no legislative gaps and no discrepancies between the various legal instruments. This approach worked well for a long period of time.

Yet the more diverse the type of data processing techniques become, the more diverse the parties that have access to the technologies and the more diverse the goals for which they are put to use, the less an omnibus regulation seems the right type of regulation. In the 1990s, there were still relatively few data processing techniques available, and there were relatively few parties with access to them. Now, not only big corporations and governmental organizations, but virtually everyone has access to advanced data processing technologies. These technologies may serve a variety of means. Medical institutions that do total genome analysis, for example, are in no way comparable to citizens that use drones and spy products; the way in which smart cities, and private-public partnerships use data analytics for nudging is in no way comparable with how companies extract information from public sector information that has been made available for re-use in aggregated form.

The more disparate the data processing landscape becomes, the more the question becomes relevant whether a sectoral approach should be considered. Such could work through several modes. Obviously, the GDPR already allows for and even encourages sectors to draw up codes of conduct, spelling out how the general rules provided in the data protection framework should be interpreted for specific contexts. Yet, very few sectors have drawn up codes of conduct so far. In addition, it can be questioned whether setting out one list of general rules set out works in the 21st century. An alternative could be to have specific legislative regimes for specific sectors. Currently, the law enforcement sector has its own regulatory regime, but in addition, it could be considered to adopt data protection regimes specifically for, inter alia, the financial sector, the medical sector and the gaming industry.

Scenario 4: revising the data protection framework, using clearly defined data categories

In the fourth scenario, the current data protection regime is regarded as in need of fundamental revision. Under this scenario, it is believed to still be possible to work with categories of data, even the current ones, but in light of the technological developments, the regulatory regime applied to them is in need of reconsideration. A number of regulatory alternatives could be considered.

First, a clear example that came up during this study is the need to regulate non-personal data. If the protective rationale is linked to the protection of societal interests and curtailing data power, most obligations and requirements from the GDPR can be applied, either in full or in limited form, to organisations that process non-personal data. For example, such a regime could look as follows:

Principles
Non-personal data shall be:

1. processed lawfully, fairly, and in a transparent manner ('lawfulness, fairness and transparency');
2. collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes ('purpose limitation');
3. adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation');
4. accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that non-personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay ('accuracy');
5. kept no longer than is necessary for the purposes for which the personal data are processed ('storage limitation');
6. processed in a manner that ensures appropriate security of the nonpersonal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures ('integrity and confidentiality')

Obligations
To the extent reasonable and proportionate, every natural and legal person processing non-personal data has to:

1. adopt a data protection policy that specifies how the rules in this Regulation shall be implemented and respected within its organisation ('data protection policy');
2. implement the policy decisions in its technical infrastructure by design or by default ('data protection by design and default');
3. maintain records specifying the data that are processed, the source of the data, the purpose for processing the data, the period for which the data are stored, the natural and legal persons with whom the data are shared, and the technical and organisational measures applied ('records of processing activities');
4. conduct a data protection impact assessment before engaging in specific processing activities, taking into account the likely effects on citizens, groups and society at large and developing strategies for mitigating those effects ('data protection impact assessment');
5. designate a data protection officer, who shall be fully independent, trained and have access to necessary resources to adequately fulfil their tasks; the data protection officer is responsible for ensuring that all relevant principles contained in this Regulation are upheld ('data protection officer'); and
6. process data transparently, meaning that the public is informed through a website of the data that are processed, the source of the data, the purpose for processing the data, the period for which the data are stored, the organisations with whom the data are shared, the technical and organisational measures applied, and any data breach having occurred ('transparency')

Only the data subject rights would be difficult to apply, although a general right to correct incorrect or outdated datasets to obtain information about data processing operations or a right to request to stop making automated decisions serve a general interest and could be attributed to persons and civil society organisations.

Second, another example that came up during this study, but was not discussed in detail, is the regulation of metadata, for which there currently does not exist a legislative regime. Given that the value of metadata has changed and that many organisations prefer gathering metadata over content communication, the regulator might consider developing a regulatory regime for metadata. The rules set out by the European Court for Human Rights could be used as a starting point.

Third, it could be considered to structure the data protection regime around the stage of data processing. Commonly, three stages are differentiated: gathering and storing data, analysing data and using data or the outcomes of data analysis. The current regulatory regime almost exclusively focuses on the moment that data are gathered and stored. Most data protection principles kick in when data are first gathered. It is at that moment when the ground for the processing must be determined and the purpose specified. Both the purpose limitation principle and storage limitation principle link back to that moment. The duration of data processing and the reasons for processing are limited to what is necessary in light of that original purpose. The data minimization and storage limitation principles relate back to the goal set out when gathering personal data. The obligation of transparency and providing information to the data subject is also principally linked to the moment that the data are first processed: information should be provided either at that moment that the data are gathered or when the data are obtained not from the data subject directly, the information has to be provided no later than a month after the data have been obtained. The moment data are gathered is also the moment that the security and confidentiality principle and the data quality principle kick in, though these requirements play a role throughout the process, and may change in time, as inter alia the techniques available for hacking evolve.

There are very few rules in the data protection framework that apply to other moments than the initial gathering and storage of personal data. There are virtually no rules on the analysis of data and no rules on the use of data, perhaps with the exception of one provision on the prohibition on automated decisions making, which may be exempted in any case with reference to a legal basis and consent, and is so limited that it plays virtually no role in practice, inter alia because the provision speaks of *solely* automated processing. This could be changed by introducing rules on the analysis or use of data. For the analysis of data, references could be made to the rules for statistical agencies discussed in this study. For example, a regulatory regime could be structured as follows:

Article - Gathering data
When data are being gathered, the following rules should be adhered to:

- data are processed lawfully, fairly and in a transparent manner in relation to the data subject;
- data are collected for specified, explicit and legitimate purposes;
- data are adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed;
- data are accurate;
- data are kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed;
- data must be stored safely and confidentially;
- data are only processed if and to the extent that at least one of the following applies:
    - the data subject has given consent to the processing of his or her personal data for one or more specific purposes;
    - processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;

- processing is necessary for compliance with a legal obligation to which the controller is subject;
- processing is necessary in order to protect the vital interests of the data subject or of another natural person;
- processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;
- processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.

Article - Analysing data

When data are analysed, the following rules should be adhered to:

1. Statistical principles
    a. Before analysing data, it must be ensured that the data are gathered in a neutral and objective manner.
    b. Data must be updated and updating data must be done in a neutral and objective manner and in accord to the original research design.
    c. Categorization of data must be done in a neutral and objective manner.
    d. Algorithms used to analyse the data must be objective and neutral.
    e. Data may only be used for the purpose for which they were gathered.
2. Transparency and oversight
    a. First, the methods of research and analysis should be recorded.
    b. Second, those methods should be made public.
    c. Third, any changes in the methods should be recorded and made public; errors and biases should be corrected and made public.
    d. Fourth, internal audits should be conducted to analyse the correctness and efficacy of the methods, both prior to, during and after the analysis of data.
    e. Fifth, external audits by experts or other organizations should be allowed and promoted – prior to, during and after the analysis of data.
3. Comparability and compatibility
    a. First, metadata on the database and analysis process should be kept.
    b. Second, gathering, classification and categorizing data should follow the rules and procedures commonly used by other organizations.
    c. Third, research methods and tools should align with those commonly used by other organizations.
    d. Fourth, there should be an equal spread in data about parts of the population.
    e. Fifth, when databases are integrated or merged, categorization and analysis should ensure the reliability of the merged data set and the data analysis following from it.

Article - Using data

....

Finally, it should be noted that these regulatory options all take the general data differentiations made under the current regulatory regime as a starting point, but add different rules to them. It could also be considered to use different data categories altogether. This study has not identified which new data categories may or should be considered in this light; moreover, such a fundamental reformulation of the data protection framework should be part of a democratic debate and extensive legislative debate. Still, as a general point, this study has found that the choice to focus on 'identifiability' may have been logical several decades ago and may have worked well for a long period of time. Currently, it is questionable

why the question whether a person is identified or not should be relevant and decisive for the application of the data protection framework. Technological experts have questioned this approach.

Scenario 5: revising the data protection framework, removing clearly defined data categories

A final scenario is one in which it is simply impossible to work with different data definitions and to attach to each of those different levels of regulatory protection. Instead, a fully contextual approach should be taken, fully dependent on a case-by-case analysis of the potential harm that results from a certain processing operation. Such harm could be linked to individual interests and/or societal interests. Most of the current obligations and requirements could be left intact, yet they would be made dependent on the level of risk and harm. As discussed, it is equally questionable whether the different categories of agents can still be upheld, because the roles of data controllers, data processors and data subjects are increasingly fluid. Leaving that discussion aside, the GDPR could essentially be boiled down to a simple set of rules along the following lines:

1. Taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of the processing, the nature of the data, as well as the risk of varying likelihood and severity for the interests of natural persons and/or of societal interests, the controller shall ensure that:

- data are processed lawfully, fairly and in a transparent manner in relation to the data subject;
- data are collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes;
- data are adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed;
- data are accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay;
- data are kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed;
- data are processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures;
- a data protection policy to the extent appropriate is adopted;
- data protection by design and by default measures are implemented;
- an assessment of the impact of the envisaged processing operations on the protection of personal data is carried out;
- the person, group or category affected in a full and detailed matter of the data processing initiative is informed;
- data are only processed if and to the extent that at least one of the following applies:
    - the data subject has given consent to the processing of his or her personal data for one or more specific purposes;
    - processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;
    - processing is necessary for compliance with a legal obligation to which the controller is subject;
    - processing is necessary in order to protect the vital interests of the data subject or of another natural person;
    - processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;

- processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.

2. Taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of the processing, the nature of the data, as well as the risk of varying likelihood and severity for the interests of natural persons and/or of societal interests, the controller and, where applicable, the processor shall ensure that:

- appropriate technical and organisational measures to ensure a level of security appropriate to the risk are implemented;
- that in case of a data breach, the relevant (joint) controller is informed, as well as, the data subject and the data protection authority;
- a record of processing activities under its responsibility is kept, which contains full and detailed information on these activities;
- a data protection officer is appointed.

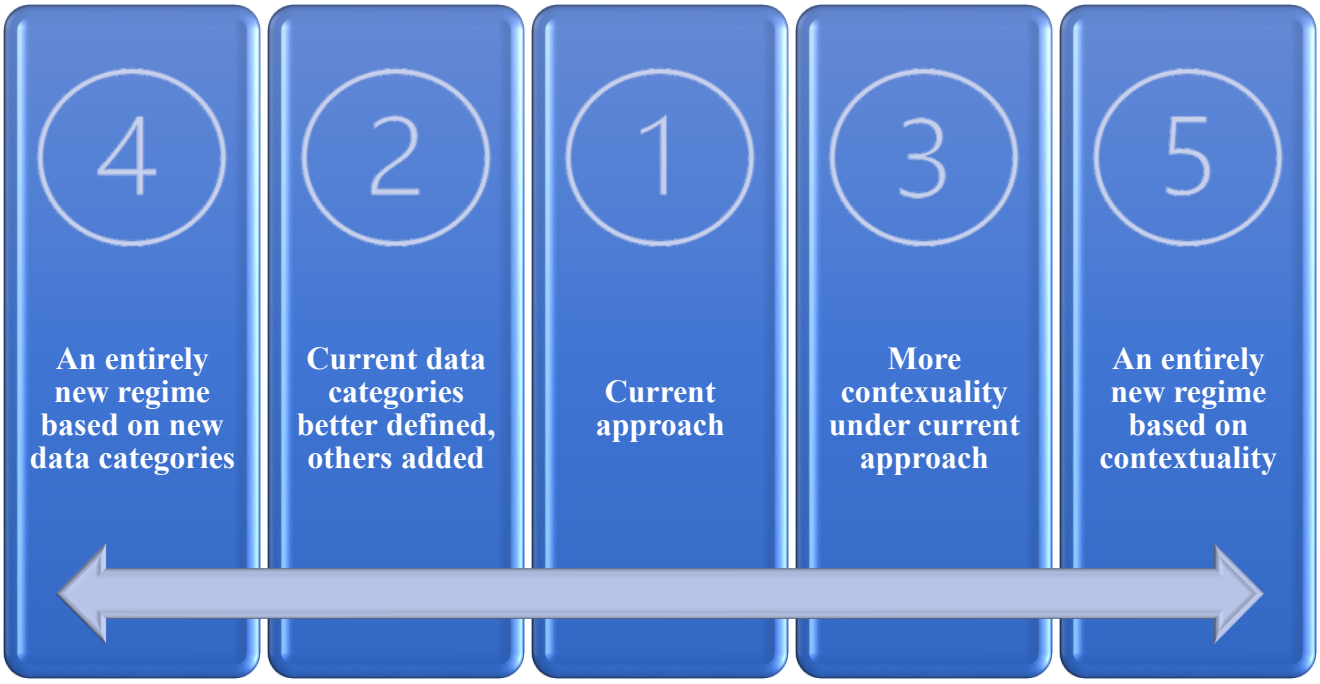

Figure 7: Scale from a fully categorical approach (option 4) to a fully contextual approach (option 5)

Finally, it should be stressed that these scenario's address the gap between the legal and the technological realm by altering the legal regime. It could be argued that the technological realm should be altered to address the gap. Without intervention, the technical experts interviewed and the literature assessed for this study are generally on par that the technological landscape will evolve in such a way that the legal data categories will be more and more difficult to uphold. An intervention could be conceivable, the most likely one being a legal one. On this point, however, legal experts are sceptical; they doubt the capacity of any legal regulation to stop or bar the larger technological and societal developments. In addition, it is underlined that Europe cannot block, stop or prohibit technologies or the datafication of society in isolation; data and data technologies will be developed and used in other jurisdictions and this will have an inevitable impact on data processing operations in Europe and data about European citizens. Though the EU can put a hold on certain AI systems or applications, such as the facial recognition by law enforcement authorities, it cannot prevent AI, Big Data and Quantum Computing from further developing and the society as a whole from becoming datafied to an even greater extent.

6.7 Answers to the research questions

1. What means are available to link (anonymous) data back to individuals and to what extent does the availability of other (e.g. open source) data play a role?

There are many means available to link data back to individuals. This study has not arrived at a full and exhaustive list of possibilities, but has discussed a number of common means for doing so. Examples are database reconstruction attacks (through which an aggregated database is re-identified), composition (through which two or more anonymized datasets merged together can result in (sensitive) personal data), and several de-anonymization technologies. Information may be inferred from anonymized datasets about people that were not in the dataset in the first place, and that aggregated data, in particular, may be used for decision-making processes which may have a significant effect on citizens in general and specific groups in particular. If the latter is the case, those data may qualify as personal data.

Open data plays an important role in this respect, so much so that many experts point out that although it may be possible to de-individualize a dataset taken in isolation, because it is possible to combine it with other data freely available online, it can never be excluded and, to the contrary, will be increasingly likely that an anonymized dataset will in time be de-anonymized by one party or another. Aggregated data, when they are made available, may be used for decision-making that affects specific identified or non-identified citizens. How data will be used cannot be controlled or estimated with certainty beforehand. However, the chance that when data are made available online, they will be used by a party in ways that have an effect on concrete individuals, groups or society at large is increasingly likely.

2. Which (technical) developments are expected in the coming years with regard to the means to (intentionally or unintentionally) link data back to persons?

It will be increasingly difficult to ensure (legal) anonymity. Already now, experts interviewed for this study doubt whether it is possible to meet the legal criteria for anonymity. While the legal regime treats anonymity as a binary matter, most technical experts see it as a scale. Most technologies and counter-technologies are involved in a never-ending cat and mouse game. Such is also believed to be the case for the future of, inter alia, anonymisation and de-anonymisation techniques, aggregation and inference techniques and for encryption and decryption. What is the most fundamental shift is the general availability of such technologies. This means, especially when data are made available online, it is increasingly likely that there will be some parties around the globe that will use advanced technologies to decrypt, re-identify or de-anonymise data and invest the necessary time, energy and effort for doing so. A potentially revolutionary technological development can come in the form of quantum computing. Quantum computing is believed to be able to break most, if not all, forms of current encryption, just like current techniques can decrypt Data Encryption Standard (DES) encrypted messages from 40 years ago.

3. What current and foreseeable technical developments can be used for the anonymisation or pseudonymisation of personal data and what factors are decisive in this respect?

Various techniques exist for both anonymisation and pseudonymisation. Examples of anonymisation techniques include, but are not limited to: masking and using synthetic data. There are various factors that are decisive, but much depends on whether a technical or a legal approach is adopted. Also, in technical literature, various types of anonymity, each with their own emphasis on different factors, have been put forward, most importantly: k-anonymity, l-diversity, t-closeness and ε-differential privacy.

For aggregation, a difference can be made between, inter alia, aggregation based on third parties, aggregation based on data perturbation, and aggregation based on cryptography. Each of those

underlines different factors that are deemed to be decisive. Perhaps the most important technique in terms of aggregating data, especially in light of data disclosure, is SDC. There is no fixed standard for SDC; each agency may adopt its own factors, standards and thresholds, taking account of the dataset, its value and potential privacy risks.

Several pseudonymisation techniques exist, most importantly, for the purposes of this study: hashing, key hashing, salt hashing, and pepper hashing. Encryption is, legally speaking, seen as a sub-set of pseudonymisation. Several encryption techniques exist, most importantly: symmetric encryption, asymmetric encryption, homomorphic encryption, and multiparty computation (which is more than merely an encryption technique). The latter is a technique that deals with protocols which allow a set of parties to jointly compute a function of their inputs or identifiers while avoiding revealing anything but the output of said function.

4. What technical developments in the area of anonymisation and pseudonymisation of personal data are to be expected in the coming years?

Most experts interviewed and the literature evaluated for this study do not expect a technological revolution in terms of anonymisation and pseudonymisation, but rather expect the cat and mouse game to continue over the coming years. However, due to the general availability of data and the general availability of technologies, it may become even harder to arrive at anonymous or pseudonymous data. Quantum Computing, as said, could have an important impact on encryption. In addition, Deep Learning is a technology that is expected to gain even more prominence over the coming years. Both technologies can have a detrimental effect on privacy, but they can also be put to its advantage. Post-quantum encryption is believed to be much safe than current forms of encryption, and deep privacy tools (privacy tools based in deep learning models) are currently being developed.

5. What can be said, from a legal and technical perspective, about the interpretation of the concept of 'means reasonably likely to be used': what means can be considered reasonably likely to be used and what factors play a role in this?

From a legal perspective, both the CJEU and the WP29 have emphasised time and again that the assessment of which means are deemed to be reasonably likely to be used should be done on a case-by-case basis, taking account of all relevant circumstances of the case and having an eye to various relevant, but not in themselves determinative factors, such as the costs of and the amount of time required for identification, the available technology at the time of the processing, and technological developments. Although these are objective criteria in and by themselves, their interpretation depends on the context. Thus, though the distinction between non-personal and personal is binary and absolute in its legal effect, the criteria to determine whether data are anonymous are highly contextual.

From a technical perspective, the contextual approach is most apparent. Most technical experts do not believe in absolute or full anonymity, but rather point to a scale of how difficult it is to de-anonymise or re-identify a database. Because technological capabilities for de-anonymisation are evolving, an assessment of the technical standards to anonymise data might need to be permanent or periodical. In this light, a black-and-white distinction between anonymous and non-anonymous data is not obvious; rather, from a technical perspective, it might be more appropriate to work with a scale under which the more anonymous data is, the less (strict) data protection standards apply. There is no exhaustive list of factors from a technological perspective that should be taken into account in order to determine the means reasonably likely (a legal notion that is not standardised in most technological discourse).

6. How does the answer to question 5 relate to developments in current and expected techniques to achieve anonymisation and pseudonymisation?

The general availability of open data and the general availability of data technologies will have a threefold impact on the possibilities of achieving anonymisation and pseudonymisation.

First, the nature of the data in Big Data processes is not stable, but volatile. A dataset containing ordinary personal data can be linked to and enriched with another dataset to derive sensitive data; the data can then be aggregated or stripped of identifiers and become non-personal, such as aggregate or anonymous data; subsequently, the data can be deanonymized or integrated into another dataset in order to create personal data again. All this can happen in a split second. The question is, therefore, whether it makes sense to work with well-defined categories if the same 'datum' or dataset can literally fall into a different category from one second to the next and into still another the next second.

Second, as a consequence of the previous, it is increasingly difficult to determine the status of data precisely. In order to determine the current status of a datum or dataset, the expected future status of the data must be taken into account. Given the general availability of technologies and the minimal investment required, it is increasingly likely that when a database is shared or otherwise made available, there will be a party who will combine it with other data, enrich it with data scraped from the internet or merge it into an existing dataset, but also that there are other parties who will not. The legal category to which the data belongs is therefore no longer a quality of the data itself, but a product of a data controller's efforts and investments. Consequently, it is arguable whether anonymization or pseudonymization can be achieved in a context where the determination of the status of data is hardly attainable.

Third, modern data processing operations are increasingly based on aggregate data, which can also have very large individual and social consequences. Profiling target groups rather than individuals is becoming a prevalent processing operation in the information society. The consequences of these activities can be negative for the group, without the damage being directly relatable to individuals. The idea that the more sensitive the data are and the more directly they can be linked to a person, the more strictly its processing should be regulated can therefore be questioned. In addition, the question is whether the focus on the identifiability of an individual (natural person) and, subsequently, the notions of anonymization and pseudonymization which are built thereon, are viable in the 21st century.

7. When is it reasonable to say that data can no longer be linked back to an individual and that the dataset of which they are part can be considered anonymous?

While, from a legal perspective, there is a difference between non-personal and personal data, from a technical perspective, this distinction falls apart into at least three relevant subcategories:

4. the situation in which data was never personal before, but might be, such as when weather data are used to make decisions about the insurance of individual farmers.
5. the situation in which data were personal, but the identifiers have been stripped or data has been render anonymous in such a manner that cannot identify the data subject nor make him or her identifiable. Here, the danger is that data are re-identified or de-anonymised.
6. the situation in which data are aggregated. Here, both the danger exists that data can be de-aggregated, that two datasets combined can yield personal data, and that aggregate data can be used to making decisions that have an impact on individual data subjects or single them out, without knowing their identity.

For each of those scenarios, different threats exist. From the technological domain, it is clear that it is almost never reasonable that data can no longer be linked back to an individual. There are always risks of de-anonymisation, there are always possibilities of data composition, and it can never be excluded

that data will be used for singling out non-identified individuals or for developing decision trees that have an impact on groups and/or individuals. As a result, it is increasingly difficult to affirm that data can no longer be linked back to an individual and that the dataset of which they are part can be considered anonymous.

8. To what extent is the test for indirect identifiability objectifiable?

Few cues have been found to make the test more objectifiable. It is important to underline that making the test objective was not the desire of the EU regulator. On the contrary, the current open, contextual, and fluid set of criteria were favoured over the more restrictive ones that were considered and rejected. For example, the initial proposal for the Data Protection Directive did not contain the notion of anonymity, but rather that of 'depersonalisation', which was understood as modifying information in such a way that it could no longer be associated with a specific individual. The explanatory memorandum provided that '[a]n item of data can be regarded as depersonalized even if it could theoretically be repersonalized with the help of disproportionate technical and financial resources'. At the same time, the explanatory memorandum defined depersonalization as 'modify[ing] personal data in such a way that the information they contain can no longer be associated with a specific individual or an individual capable of being determined except at the price of an excessive effort.'[287] Excessive effort is still contextual, but less so than 'all means reasonably likely'; also, the threshold is clearly different.

Few cues have been found in this study for making the test of indirect identifiability more objective other than deleting the notion of 'identifiability', which was not originally part of the definition of personal data under the data protection regimes from before 1995, or limiting the list of factors to be included for determining what means should be deemed reasonably likely to be used. Perhaps the only concrete suggestion that was identified is putting a time limitation or a horizon to the evaluation of the means reasonably likely to be used. It is almost always highly likely that, in 20 years time, data that are anonymous now can be de-anonymised. Under the current legal regime, when data are stored for that long, such means reasonably likely to be used must be taken into account when determining whether the data protection regime applies, while it is next to impossible to foresee how the technological landscape and the availability of data will evolve in the next 20 years.

9. To what extent and in which cases can there be underregulation when data are no longer linked to individuals through anonymisation and therefore do not fall within the scope of the GDPR?

10. To what extent and in which cases can there be overregulation when more and more data can be easily linked to individuals through new techniques (undoing measures of anonymisation and pseudonymisation)?

Answering questions 9 and 10 depends on what is deemed to be the regulatory objective of the data protection regime: is the data protection framework to be considered from a protective angle or from the perspective of facilitating data processing within a set framework, or as a combination between both? Is the protective rationale to be understood as primarily providing protection to individual interests or to group and societal interests? Should the data protection regime be understood as laying down limitations for data processing or as providing a framework for using and sharing data? Is the protective rationale best served by limitations, or can more data processing sometimes be required to serve the best interests of individuals and/or society? Is the rationale of facilitating data use best served by an open and contextual framework or by setting strict and clear rules within which data processing is deemed legitimate? This study has not been able to give a determinative answer to these questions,

[287] COM(90) 314 final ~.sYN 287 and 288 Brussels, 13 September 1990.

but has indicated that dependent on these answers, different regulatory gaps and dangers for over- and/or under-regulation will be found.

For example, whether there is underregulation because 'personal data' is linked only to the identifiability of natural persons and because the data protection framework refers primarily to the interests of the data subject depends on which rationale the data protection framework is said to protect. If it is considered that the data protection framework is or should be providing protection to more general, group or societal interests, then certainly, there may be a matter of underregulation due to the fact that processing aggregate and anonymous data is not covered under the current regime. Likewise, whether the trend of courts and advisory bodies to expand the scope of personal data and the material scope of the data protection framework leads to overregulation is dependent on whether the emphasis is placed on the protective rationale of the data protection framework, in which case there would be no overregulation, but to the contrary, this approach could be deemed laudable, or on the rationale facilitating data processing, in which case it may be deemed stifling.

11. How will the current and future technical developments affect the GDPR and legal protection in a broad sense in the coming period?

It is clear that the technological developments and general availability of data now and in the future have the effect that anonymisation will become increasingly difficult. The status of data will become increasingly volatile and will be less and less a characteristic of data and datasets themselves and more and more an effect of the data controller's efforts. The legal categories will become more and more fluid and porous, and one database may be legally qualified differently per party that has access to it. A database that in isolation only contains non-personal data may be turned into personal data by combining it with another database the next moment, may be used to infer sensitive personal data the next, only to be aggregated and anonymised the next moment again. Given these trends and given the notions of 'identifiability' and 'all the means reasonably likely to be used', more and more data, if not all, will fall under the data protection framework.

This study did not find different scenarios for how the technological realm and the availability of open data will develop over time - literature, experts interviewed, and experts invited to the workshop held for this study all point in the same direction. Several scenarios were found, however, for how the legal regime could respond to the increased availability of open data and the general availability of technology. Five strategies were deduced from the suggestions: leaving the current data protection framework intact, focussing on clearer data categories, focussing more on contextuality, using different data categories and regulatory regimes attached to them, or focussing on a full-blown contextual data protection framework.

Chapter 7: Annexes

WODC supervision committee

Prior to the research, the WODC set up a supervisory committee. The members of the supervisory committee have personally supervised the quality and independence of the investigation and have supervised that the investigation is consistent with the initial proposed research. The advisory committee consisted of:

- prof. dr. N. Helberger, Universiteit van Amsterdam (head of the committee)
- dr. F. Dechesne, Universiteit Leiden
- dr. J.H. Hoepman, Radboud Universiteit Nijmegen
- mr. dr. M.H. Paapst, Rijksuniversiteit Groningen
- D.D. van der Neut LLM MSc, Ministerie van Justitie en Veiligheid
- dr. L.M. van der Knaap, Wetenschappelijk Onderzoek- en Documentatiecentrum

7.1 Interview reports

7.1.1 Bruegger & Hansen

Interview 12-05-2022
B. P. Bruegger[288] & M. Hansen (Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein)

Question 1 – How does the GDPR deal with knowledge and more specifically meta-information? It seems like the GDPR is more focused on data. The idea of information is perhaps included in art. 4 GDPR on identification, but are people aware of that?

- Meta-information is often overlooked by technical people. There is perhaps some prototypical thinking underlying the GDPR, so the GDPR mentions information but they are mainly thinking of data. If we for example take the question when is someone identified, sometimes it is knowledge or information in your head that allows that identification. I try to retrofit that into the wording of the GDPR. Recitals and provisions are often prototypical, so there are drafted with a specific situation in mind, but the situation does not always apply.

Follow up question: it is interesting that you phrase it as a misconception, and that often the GDPR refers to the term information while in reality it concerns data. How do we deal with misconception?

- There are some real-world measures that take that into account, but normally it is not acknowledged that it is also about information or knowledge in the head.
- But of course the GDPR is not so much about information in the head but rather about the identifiability question. For example when we talk about an organization, what would the 'information in the head be there', maybe that concerns employees. The first step is what data is regulated in the GDPR and the answer is that it is only the file system of information, it is not about information in your head. But identifiability applies not only to information in the file system, however in practice you need to have proof that the information was there if you want to investigate as a DPA. Thus, the information needs to be embodied. Meta-information is usually not all in your head. So, the example of information in your head is a good example, but does not fall within the GDPR. It is not excluded for the future, for example in the meta verse, it is a different discussion.
- Maybe we need something that is an in-between the head and the in file. De-identification attacks often rely on meta-information, it can materialize in the head (information that was gathered from various sources or was guessed). So, the information can be found all over in different places, but the de-identification idea materializes in the head.

Question 2 – Can you explain more what you mean by the identity domain (regarding unique handles and non-unique handles)?

- In the identity domain you get assigned an identifier, some kind of a string or a number. E.g. a passport number or social security number given to you by the relevant authorities, or a nickname in a social group that was given by friends. So someone assigns the identifier.
- The assignment is typically unique in that context. For example: in an organization you can have IP addresses that cannot be used outside of the organization, outside of that context it is meaningless.
- Identification happens in a certain context, and you need that domain knowledge. E.g. you can be given an Italian and an American social security number, but you need the domain knowledge to be able to identify those individuals. So normally someone would think just because you have a social security number you can identify that person, but for that context it might not be true.

[288] The reported work referred to in this interview is part of the projects PANELFIT (https://panelfit.eu) and TRAPEZE (https://trapeze-project.eu/). Both have received funding under the European Union's H2020 research and innovation programme under grant agreement numbers 788039 and 883464, respectively.

Question 3 – In your publication you mention different types of linking (and also special types of linking e.g. linking based on an implicit order in the data, model-based linking, attribute linking). Which type is most worrisome for the data subject or from the GDPR perspective? Or does the type of linking not matter so much in that sense?

- The GDPR only thinks about certain cases, so I certainly know this is a specific person. While speaking more broadly you could think of a situation where you have a certain chance that it is a specific person, e.g. I know one out of three certain people has a specific medial condition, I don't know exactly who it is out of the three but there is a 1/3 chance. The GDPR has this prototypical thinking, does a certain situation apply yes/no. With AI or neural networks these discussions come up more, as you can link between correlations which might be aspects that are seemingly harmless/less relevant, so we might be less aware of the actual dangers. While with the use of AI and neural networks in combination with the data that is out there, we might be able to link more and more. Is it by law counting as identified if you do not know how certain it is, perhaps not, but the effect on people could be there. The GDPR does not use the term linking, the German standard data protection model has the term linking. GDPR does not identify the process how to identify (does not mention linking or establishing a link). The law of course cannot stipulate the technology that can be used, but it assumes some sort of correlation.

Question 4 – On the topic of pseudonymization: you could take different perspectives, one could advocate that the GDPR is about preventing identification as such (that it would be harmful as such), or you could propose that it is not so much about identification as such but rather about thinking whether we need to identify in a specific context. So could we say in itself it is not that much a problem that individuals are identified, rather the question is if it is always necessary. That is why we have anonymization (if identifiability is not necessary) or why we apply pseudonymization when possible or necessary. Can you explain how you view the concept of pseudonymization?

- The GDPR assumes that if you don't need data you do not store it, this comes with data minimization. Data minimization is not just about the volume of data but also about the types of information. Thus, you pseudonymize or anonymize data if you can, if you do not need that information. So you have that risk factor, how much data do you have and how rich are those data. The second risk factor is how strong is the link to the person. For anonymous data the GDPR assumes that it is out of the risk zone, pseudonymization is the middle ground. For pseudonymization, in a certain context you prevent that certain people can be identified, it reduces the risk. The GDPR is quite clear on that. However, everything can fail, so pseudonymization is in a way also a second line of defense.
- Data minimization is important but also accuracy for example is important, especially with data processing on a big scale. So other principles apply as well, we do not only have pseudonymization or data minimization. Still, pseudonymization could be used much more. The same goes for encryption. However for both we can say it can be good to use but it is not always necessary to do it whenever it is possible, it depends.
- Why the confusion: there is data minimization and storage limitation. Data minimization already requires storage limitation. Perhaps we should have used the term identification minimization rather, to detangle the storage limitation from data minimization, then it would have been clearer whether you for example have to pseudonymize or not.

Question 5 – In our previous ULD interview at the end, it was mentioned that the GDPR is not perfectly technology neutral as it can favor some PET's over others. What do you mean by that?

- Usually there is one controller that accesses the personal information and is responsible, what about those ideas and mechanisms that rely on de-linking information. So, there can be a situation where one controller may have the personal data and other processors or controllers might have parts of the rest. Only if all the information is put together you can see the full picture. This means who knows now whether they are part of a system with personal data?

They have no chance to know that. Only if someone has all the mechanisms to join that information do they know what is the bigger picture. So, if we rely on mechanisms that divide the data so that each controller may not misuse the data, it can be hard to know for that what their responsibility is. The law wants to address one controller, or a joint controllership. So those who are offering something else might be discriminated. The GDPR can only certify controllers as well, not other actors. So you cannot be certified if you do not process personal data, because you first need to be a controller.

- The GDPR is focused on one controller, the real world is not. Tools such as data protection impact assessment are difficult to apply, as the privacy is a component of the system, not of the part run by a single controller. E.g. ISP's do not know who someone is but what they do on the platform, while identity providers know who someone is but not what they do. This is a distinction. We have 'initial controllers' but they can also sell the data to a next controller. Data subject rights should apply not only to the 1st controller but to everyone who uses the data. The prototypical thinking of the GDPR was to have one controller. But if you have a chain of actors, you might want a dashboard etc., this is not strongly covered by the GDPR.

Question 6 – It is sometimes proposed that it is very difficult to achieve true anonymization of data, or that if data are really anonymized they might not be so useful anymore. In the guidance you sketch on the one hand the scenario of damage control with data that were falsely presumed anonymous and on the other hand the scenario of playing it safe. What does that mean for the concept of anonymous data, as there is a risk of not providing enough protection, or 'over doing it'. Would it have been better to have a different approach or concept than the current one?

- It is a success state, but no one can ever be successful in that regard. The wording of the GDPR is without a time horizon and also future technology and including any additional information so also information that you might not have. In that approach there is no way of telling whether something is anonymous or not. There are good uses where you want to work outside of the GDPR, e.g. in medical research or road traffic control. There is no way of seeing when it would be enough. If there is still a risk of re-identification, there is a risk of harm. Thus, data controllers want to know when it is enough and when it is not enough. But you cannot answer that. It is not a solvable problem to some extent. Maybe we can say all data is somehow personal data, but that can also hamper initiatives that are good for society. (e.g. data markets and data commons)
- We also see initiatives with the data act and other instruments covering anonymous data. It is difficult, there will be error cases and court cases. There might be additional guidelines or best practices etc., then it will be clear what to expect. There might be clear cut cases but not always. Re-identification could be forbidden, but that is not a good idea, to provide only a legal solution. There are no ways from a computer science perspective to say whether we have anonymous data in this semantic world. So better to apply the GDPR whenever possible. Being careless or not is also an aspect that we take into account.
- Big danger in policy making: EU policies focus on the value of data sharing and data markets, but then what they propose does not comply with the GDPR. We have to be honest about compromises, it is not easy to do data markets or commons that are GDPR compliant. We need more guidance on this.

7.1.2 Drogkaris & Bourka

Interview 10-05-2022
P. Drogkaris - Cybersecurity Expert & A. Bourka - Data Protection Officer (ENISA)

Question 1 – Currently there is so much data out there. From your point of view, does this create a challenge to maintain anonymous data and pseudonymous techniques? Or is there for example also a lot of development in terms of privacy engineering to equally increase the strength of pseudonymization techniques? (Can you describe the situation/state of the art)?

- This is a challenge already for several years, we also outline it in our reports. It was already mentioned in the 1st report of the Article 29 Working Party on anonymization. We have a chicken and egg situation between technologies and new technologies to bypass that protection. This is why we have such a broad definition of personal data. And this is also why we have no specific definition of anonymous data. Because the means change all the time. In our reports we focus more on pseudonymization techniques then anonymization, and of course it is important to keep in mind that the two are completely different in scope. Pseudonymization is not intended to make data anonymous, it is supposed to protect the data as a security measure, it should be possible for example for the controller to go back to the data. Protection is a continuous process; we always need to find new ways of protection as new technologies emerge.
- We should be careful when discussing whether data are anonymous or pseudonymous, as the two are completely different things. It is also important to remember that there are other legal instruments apart from the GDPR that refer to anonymous data and to de-anonymization. Thus, the EU legislators also keep in mind that anonymous data might not stay anonymous forever, and this is also in line with what the working party put forward back then about anonymization. The technology progresses so we have new opportunities and new interconnections are possible between datasets, given also the broad definition of personal data and the scale of the data-sets it is now easier to possibly identify an individual. Thus, the evolution of technologies presents new opportunities and at the same time challenges for personal data. The legal provisions are there, but sometimes entities do not engineer the technologies in practice perfectly, this is where data protection engineering comes into play.

Question 2 – So we do not have a definition of anonymous data in the GDPR. Do you think that has a positive or negative impact? Should we advocate for a specific definition, or keep the current approach where we define personal data and what does not fall within that definition is not personal data?

- This has been a highly debated topic previously. Personally, I think we don't need a definition exactly because we have this fluid situation regarding what is personal data and what is not. More informational also falls within the scope of personal data: identification can be based on many other data that we previously did not perceive as personal data, but they become personal data in a context with other data. Thus, this leaves the scope of personal data quite broad but also allows us to address new challenges. It is hard to speak of truly anonymous data, perhaps we can speak of aggregated data, but truly anonymous data is hard to achieve. The current approach is best because it gives us the opportunity to adjust and be flexible in protection of personal data.

Question 3 – It is hard to achieve true anonymization these days: do you see this as problematic? Or is it perhaps not so problematic that we can de-anonymize if we consider that if data are not truly anonymous, they will receive the protection of the GDPR?

- It is not so much a question of if we have the right technologies, we have some adequate technologies or techniques to anonymize a dataset. The problem is what can happen with the dataset afterwards, e.g. perhaps other anonymous data-sets can be combined with that data-set and that might lead to re-identification afterwards. Thus, after the anonymous dataset is released, we don't know what will happen to it. One of our recommendations in the latest ENISA report

is also exactly that: we give an example of an anonymized dataset, but we also make clear that you can never be sure whether this will remain anonymized once it is published.

Question 4 – We do not know which information will have an adversary that can re-identify, it is impossible to know. Based on that, do you think we should change our approach and have some probabilistic scenarios when releasing data to see which measures are likely necessary to presume the data to be anonymous data?

- In a certain way we could say anonymization is a risk-based approach, however I would take step back and question what the need for the anonymous data in the first place is; what can the anonymous dataset offer that cannot be offered by something else. I'm not talking about statistics here. E.g. the 2007 Netflix case: Why release it as anonymous data in the first place and why not treat it as personal data?
- I would stress that it is not a matter of anonymization or pseudonymization, the purpose of these two techniques is different. Anonymization is applied when we do not need personal data, while pseudonymization is a security measure, a protection of confidentiality (e.g. see article 32 GDPR). It is not a choice between the two, the purpose of pseudonymous data is also for the data controller to go back to the original data. There is this trend to say, 'if I cannot anonymize the data I will pseudonymize the data', while that is very risky given that pseudonymous data has this particular purpose and is thus very different. Of course, pseudonymous data are personal data.

Question 5 – Do you think the GDPR offers enough incentives for controllers to apply pseudonymization as a protective measure?

- Yes, the GDPR offers them. For example, the use of data for research, e.g. in the medical sector. More broadly speaking, the GDPR follows the risk-based approach, thus we have different types of levels of obligations. You can see this in impact assessments but also in the transfers of personal data. Pseudonymization is a strong protection measure if done properly, as pseudonymization measures protect the identification data, which is the key factor. Thus, pseudonymization can assist the controller to do things that otherwise would not be possible.
- We cannot expect from a legal instrument the role to provide concrete guidance. We need concrete guidance on pseudonymization and encryption. Originally encryption was perceived as a solution to a lot of problems, but we have to nuance this a bit: in practice we have to deal with questions such as who handles the encryption keys, how good are the encryption keys and the encryption algorithms. It is not a matter of choosing the right technology but rather the whole process, i.e., including the design, understanding what we need, what technologies can offer, if we need a combination of techniques, and how to make them work for a specific processing operation. But we cannot have guidance for every processing operation imaginable. The GDPR takes a step in the right direction by requiring impact assessments in some cases, so that reflection on the process is needed.

Question 6 - Continuing on the topic of guidance: at the end of the March 2022 report from ENISA on pseudonymization techniques, it is stated that 'Developers and regulators at the national and European level should promote the exchange of good practices and provide practical guidance on deploying pseudonymization in practice.' What should such a guidance include in your opinion?

- The idea is that somewhere out there, there is already the knowledge of how pseudonymization techniques can work in practice and increase the level of protection in a specific processing operation. Before we could have guidance or technical standards, we first need to have that discussion on good practices on what can be a good way forward.

Question 7 – We have the notion of identifiability in the GDPR which is related to the risk of re-identification. The Article 29 Working Party also speaks of linkage attacks. In this context, according

to the ENISA reports, there are two important points that should be considered while evaluating any pseudonymization and encryption technique: whether third parties can reproduce the pseudonyms that a data controller creates across domains; and whether the pseudonyms used can be easily re-identified. Could we say that pseudonymization in that sense focuses on identification and linkage, and perhaps anonymization is more related to information inference? Or what is the background of those two assumptions for evaluating pseudonymization techniques?

- I would re-iterate that there is no such direct distinction between anonymous and pseudonymous data, selecting and performing pseudonymization is of course completely different from anonymization. You cannot go from one to the other, pseudonymization and anonymization are different. Regarding the risk of re-identifiability with pseudonymous data, additional information provides an opportunity for the pseudonymization to be broken. Pseudonymization is not the same as encryption, for pseudonymization the recipient of the data cannot go back to the main text or the original data, only with the use of additional information this might be the case.
- Re-identification or linkage is not only about one pseudonym, but there could also be other additional information or there could be multiple pseudonyms. Other data does not even have to be identifiable information, just having additional information could be enough. This is where inference comes in, it is not just about linkage. Inference can happen in so many different levels, we can have datasets where we don't anticipate personal data. Also, by combining anonymous datasets you can create personal data. Thus, inference is also related to anonymous data, but not only of course.

Question 8 - If you can infer information, there is also the risk that you infer sensitive information. Thus, it can be difficult to maintain the strict protection for sensitive information, if we also have proxies or other additional information. Is that also a challenge that you see?

- This is a more theoretical discussion; the question is how this applies in practice. Again, it is important to seek the purpose, for example of anonymization or pseudonymization. If our purpose is to conduct research and we need personal data, for example for medical purposes, it does not make sense to try and achieve anonymity but rather we should treat the data as personal data and provide strong protection. Strong protection is not just strong technical measures but also includes the environment, procedures, organizational measures, etc.

Question 9 – So we have the distinction between anonymous and personal data, and within personal data pseudonymization is a technical protective measure. However, there seems to be a debate on keyed hashing and symmetric encryption where the EDPB and AEPD think that when the secret key of keyed hashing is deleted, that technique could be considered as an anonymization technique. Do you agree on that? Should the difference between anonymization and pseudonymization rest in the deletion or not of keys?

- Again, you are describing a technical measure, but we need to look at the purpose. If you delete information such as the key, that means that you don't need it. In that case you don't need to have personal data and this technique becomes an anonymization technique. In many other instances you would need this key, as a controller you need to be able to go back to the data and identify when necessary. It is not about the transition of the technology from the one to the other, it is about the use or purpose.
- This is one of the tricky parts of issuing guidance, it cannot cover everything. It cannot be said that easily if you use certain techniques that is anonymization, it depends. One can only provide example, we need good practices. For example, a good practice is that you can hash for pseudonymization, but you should also apply some additional measures.

Question 10 – There are also other means of technical protection (against identification and information inference) than pseudonymization. Do you think it is good that there is so much attention for, or emphasis on, pseudonymization, or should we look more at other measures as well?

- We see in the GDPR, for example in article 32 GDPR, that pseudonymization is one of the possible techniques. As goes for all security measures, this measure can offer a certain type of protection. There is not one measure that can tackle all problems. That is why we have risk management approach towards information security under the GDPR. So, we need different measures in different scenarios. Pseudonymization is a powerful measure, because it protects identity, which is of key importance, but not always enough in itself. Also, pseudonymization is not always possible, we need other measures as well. But pseudonymization is also not a difficult measure to apply in simple scenarios.
- Pseudonymization is mentioned in the GDPR itself as well a couple of times, so we could say that says something about the usefulness of this technique. However, pseudonymization is an umbrella for different techniques (hashing, random generators, etc.), similar to encryption. Article 32 GDPR mentions technical and organizational measures, organizational measures can also play a critical role towards compliance or protection. It should be a combination of these two, it depends on the context. It depends on the end goal, what you want to achieve. Not every controller or processor is the same as well, there can be limitations to understanding the technology for example.

Question 11 – Looking towards the future: what role will synthetic data play? And what role will quantum computing play in on the one hand protecting privacy and on the other hand breaching protective measures? Do you see a role for such technologies in the field of privacy and data protection?

- Yes, synthetic data can play a role, for example using synthetic data to develop a system. But, how well prepared are we to generate synthetic data with the same quality as real data. It could be a solution to some problems, but it has to be used in the right way. So again, we could benefit from good practices, guidance and recommendations.
- For every new technology, whether it is synthetic data, quantum computing, etc., we in the data protection community tend to see the data protection dimension. But we should not forget the implementation perspective as well, we also need usability. Synthetic data cannot be used everywhere and with the same results for example. (ENISA also addressed synthetic data in a report.) We need to achieve the purpose for which the data is being processed too.
- In the cybersecurity field quantum computing is expected to change the landscape a bit, it can say something about how robust can existing techniques are. So, it could affect existing algorithms and techniques, and perhaps it could also influence the possibilities for existing datasets to combine and find correlations. Overall, it is not just one technology, but the evolution of technology changes the landscape.

7.1.3 Hansen

Interview 02-05-2022
M. Hansen – Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein (ULD)

Question 1 – Some people propose that because there is so much data out there nowadays this also facilitates linking data and, in that way, identify individuals. Is that also a challenge that you see from your perspective?

- Yes, it is increasingly easier to link, we described this problem already about a decade ago. Think of linkage enabling data, it is typical that there will be some form of identification possible afterwards. For some types of data that linking will be easier than for others, for example with location data (telemetric research is e.g. also possible on these data). Already with reduced information you can derive a lot, think also of examples such as IP addresses. The GDPR improved on the data protection directive in that sense, as under the directive regime we had discussions for years whether cookies or IP addresses are personal data. Under the GDPR it is clearer what is identifiable, compared to the data protection directive, and the GDPR does not only concern unique identifiers but also the combinations and additional information. We do not just have concepts in data such as whether something is personal data or not, or whether it is pseudonymized or not, but we also have specific provisions on other specific types of information. For example, we have article 87 GDPR on national identification numbers and safeguards for processing those, which safeguards are differently understood by the Member States.

Question 2 – What are the challenges with the concept of anonymous data? There are debates on the concept as to whether it is truly possible to speak of anonymized data, and whether we should adhere to a strict approach to anonymization or not. What are your ideas on that?

- I would not question the concept and of course have to interpret it as it is there, which is a strict interpretation. If we follow the Article 29 Working Party Opinion, there are three ways to attack or re-identify anonymous data: Singling out, linkage, and inference. There is truly anonymous data, for example in statistics, you can use them to identify trends etc., there is scrambled data, and so on. So the question is how do you make sure for the data that you are analyzing for your purpose there is no linkage, singling out or inference possible. In practice some people/actors do not want to do that, because it is only for one research question, activity, etc., and thus a lot of work to do for every research question that you want to use the data for. Thus when I refer to anonymity I refer to it in this way, in practice most regulators or other actors simply think of leaving out a name/identifier, applying some restrictions to the data, do the risk analysis, etc. Maybe the reality forces us to take more of this latter, less strict approach, but then technically it is personal data and thus you should comply with the GDPR. It is good to reduce the risk, but then you still need to look at the safeguards and obligations from the GDPR. So of course it is difficult to have truly anonymous data, but they do exist.

Question 3 – Should we distinguish between different types of anonymous data: for example on the one hand have aggregated data/statistics and on the other hand anonymity in more micro level data?

- It makes sense to distinguish on a practical level. In the paper that we shared we also use the concept of assumed anonymous data. Because often people claim that it is anonymous data, but it is not always true and sometimes there is still a risk. For example, often people remove outliers in the data to anonymize. First of all then you need some information to assess what is done to reach anonymity, so for example how many percent is removed. Second, then you run the risk that the data is biased or excludes certain people. And if the outliers are the most interesting part of the data this makes the dataset less usable. Or perhaps you use synthetic data, so you mix data with 'fake data'. And thus, we need something to express that, to express what is done with the data, which is related to the risk. To distinguish between types of anonymous data (aggregated

vs micro level e.g.) does not work well. We need characteristics which say something about the risks/safeguards/utility but are not spoiling the definition of anonymization. We don't have a good term right now for the other data, which are not truly anonymous. There has to be an incentive to do what you can, also when the GDPR is applicable. Of course we have article 25 GDPR (data protection by design and default), but were not fully there yet.

Question 4 – There is criticism on article 25 of the GDPR for being too vague. What do you think about that?

- Yes, data protection by design can be quite vague. If you just have the principles of the GDPR and article 25 GDPR it seems clear cut, you apply the data protection principles etc. But the problem is nobody really knows how to apply the principles (e.g. what is fair treatment and what is transparent). Thus article 25 GDPR itself is hard to enforce. Article 25 GDPR only binds the controllers, in doing so it excludes manufactures, so manufacturers do not have to think about this in the design. This means that we need best practices and to educate. And paragraph 2 of article 25 GDPR, data protection by default, is even more difficult, as some companies do not want to choose the default for example because their competitors don't.

Question 5 – The Article 29 Working Party proposed several techniques, but can we still see those techniques of privacy engineering as adequate today? Do we need combinations of techniques, or do we have other state of the art techniques now?

- Already in the Article 29 Working Party opinion you already see you need combinations of techniques, there is not one technology that covers all three types of attacks for example. E.g. differential privacy even gets a 'maybe' in one category. What is sufficient also depends so much on the context. If you combine technologies, the result can also be worse or better. You can see this for example from a security perspective: in some crypto fields there can be manipulation, you don't know if the information was tempered with or not. And sometimes you only need basic, not extremely reliable component, but in combining them you cut off secret ways in. Thus, you can have guarantees from a combination of unreliable technologies, but you can also have reliable components and if you combine those, they are not reliable anymore. It all depends on the data and context you have, so there is no way to say what are good techniques or combinations as such.

Question 6 – What are the challenges that you see from your perspective in controllers using pseudonymization?

- Encryption and pseudonymization are referred to in many places in the GDPR. Pseudonymized data are mainly used for research purposes or statistics for example in health data. Sometimes one cannot guarantee what the GDPR definition requires, but they did protect the data a lot or made a big effort. But some controllers don't even do that, they merely argue that they need the data and if they don't need it anymore they will delete the data. They rather need encryption than pseudonymized data. Sometimes encryption and pseudonymization are mentioned together but obviously they are not the same. And encryption is clearer than pseudonymization. We have debates on what type of information you need, for example sometimes you don't need to know the names of people in the data or other direct identifiers. In machine learning synthetic data will come into play but also pseudonymization.

Question 7 – Obviously it is important that the GDPR acknowledges pseudonymization, but at the same time we can wonder why the GDPR acknowledges pseudonymization and not other means of protection or security in the same way. What are your thoughts on that?

- It can be something on identification where the risk needs to be addressed specifically. There are other ways to blur the identifiability but do not fall within the strict definition of pseudonymization, and sometimes those other ways might be even better. For example, if you try to do something against singling out. Non perfect anonymization (of course not in a sense to

try and escape the GDPR) could be better than pseudonymization, in terms of risk reduction. Under article 6(4) GDPR we have compatible purposes where pseudonymization is mentioned explicitly, but this is not always the case of course. Perhaps you could also derive the privilege for risk reduced information in other areas. But encryption and pseudonymization have a different standing yes. We also have the crowd effects, where decentralization is being handled, but it is not explicitly addressed by the GDPR. There might be a joint responsibility of natural persons. The technological neutrality of the GDPR is not so perfect in that sense, it can discriminate PETs because some architectural assumptions are also wired in the GDPR.

7.1.4 Jensen

Interview 03-05-2022
M. Jensen – Kiel University of Applied Sciences

Question 1 – Do you see in your line of research the impact of more data being available and as a consequence making it easier to identify individuals? I.e. do you see an effect along those lines of more data being available?

- It is indeed really a challenge, we can link this to the core challenge of data protection itself, of how to protect the information contained in data. For example, from the data point of an email address you can infer a lot of different information and in different contexts. Some of these contexts might be sensitive by nature. It is a challenge to reach a level with data that is both useful data but also no longer contains links to other contexts. That is where pseudonymization and anonymization play a role. There is a lot of misconception on the terminology, e.g. some actors claim to anonymize data but in reality they are deleting some factors and are just pseudonymizing data in some way. It depends on how much effort you put into applying anonymization and how much effort you put into breaking the anonymization. Anonymity in the sense of data being not able to be linked anymore in any context to any specific human being, that this an information theory concept. It is however hard to apply in the real world, it is almost impossible to truly have anonymous data. We could say that the GDPR concept of anonymity is a little bit different: it is about what is reasonable. What does the term reasonable mean though? It is not a precise term compared to the information theory concept. This exact boundary of when it is enough to be under the GDPR definition is under debate. This boundary is crucial for companies, as when they process anonymous data rather than personal data, as then they do not have to comply with the conditions of the GDPR. For a lot of types of data anonymization does not work, for example large scale data such as video feeds, sound files or pictures, without losing the utility. It is difficult to maintain the balance between keeping the utility that you want and reducing linkability to the maximum. One technique that you can for example use is differential privacy, where you can reduce the linkability below some threshold, but completely getting rid of it is complicated. It is rather about minimizing the risks of re-identification and misuse.

Question 2 – In your 2018 paper on Big Data and the GDPR, you mention quantifying the rate of anonymity. At the same time the law (GDPR) has a black or white approach, either data are anonymous or not. Can you say a bit more about the rate of anonymity? Do you view anonymity from a technical point of view more as a scale/rate rather than a yes or no question?

- We have grey scales in between black or white. For example, we have the k-anonymity approach. That is one way to some extent measure or quantify anonymity. Of course an approach such as k-anonymity also has its issues. The core problem is that anonymity in its binary version does not look at the context and involved entities. For example, the question is whether I can re-identify myself, as I have all types of background information and know which identification factors I might have. That is one type of anonymity that is very hard to address, as you have all the background information. On the other hand, other entities might have access to techniques that you yourself do not have. For example, machine learning or huge processing power that allow for linkage. Some AI can thus introduce new attack techniques, it is a challenge to see what those techniques can do. The important question is who is able to de-anonymize, based on what background information they already have. The binary approach of the GDPR is not perfect in that sense.

Question 3 – Could you explain a bit more about the differences between information theory assumptions and assumptions under the GDPR?

- Information theory anonymity says that there is no way to link data back to an individual or learn something about that individual (e.g. classifying the individual). So even if there is no link

to that person (you cannot say who the individual is) but you know they fall in a certain category (e.g. being male, living in a specific country) you also derive information without knowing the identity/you can still find attributes. In information theory to speak of anonymity you also should not be able find attributes of the person ore identify the person. In pseudonymization you want to keep some utility, you want to still link some data for example between datasets, but without being able to link it to the person behind the data. These days what we try to do is to identify people because then it is easy to link to arbitrary contexts. The former is linking without knowing who the person is, in information theory that still gives you some information about the person, while the GDPR speaks of (re)identifying the person.

Question 4 – Continuing on that, we could say that the GDPR is more concerned with an ex-ante approach or first step of whether you can first identify an individual. While information theory is more concerned with deriving information. Do you think we should work towards more of an ex-post control (e.g. information inference)?

- GDPR is about linking to individuals, that is less strict than information theory, because then you should also not know attributes. If you know a certain attribute, such as knowing an individual is a smoker, you exclude all the non-smokers, thus it is a factor that limits the identity pool. Information theory would not say this is anonymized, while under the GDPR approach if you can for example bring it down to possible individuals it is still anonymized. Of course, the former is more a theoretical approach. In reality there are so many contexts and so many ways to link back to individuals it is almost impossible to have a system that has complete information theory anonymity. Also good to keep in mind is that data is the information carrier, it is not information itself. For example, data can be '3' but that does not give you any information on what the number 3 represents. If we strive for a system with an information theory level of anonymity, such a system would have little utility. Most systems are human made or human impacted, so there is always a link to humans. Sometimes you can link but it does not cause a privacy infringement. The question is whether the maybe not so relevant information that you leak is critical, should the GDPR concern itself with that. The big challenge is that we don't know what data leaks what information. For example, machine learning introduced lots of new ways to link data that we did not anticipate. Probably there will be more new ways to attack data or create linkage in information in the future, we have already gathered/stored so much data (big data hype). This is where the idea of anonymity also comes in, that we want to prevent ex-post data analytics on that data that would re-identify people, while we have not yet learnt to perfectly anonymize. This is partially because new technologies pop up, this is partially addressed in the GDPR in the impact assessment in case you use novel technologies, but in terms of anonymization it remains difficult.

Question 5 – You discuss how information theory is more targeted at attributes compared to the GDPR which focuses more on identity. On the one hand the former could offer a stronger protection but is perhaps not completely feasible in the real world. To what extent do you think the law can still learn from information theory?

- Information theory anonymity makes data useless for most context, we need to retain some utility. We just have to know that every classifier or attribute discriminator that we learn can be used for good or bad. The problem is that with the anonymity approach we forbid linking back, which is not very feasible. Perhaps it is better to focus on the control of the process instead, e.g. for which use is data re-identified. However, because of the different purposes and contexts possible that is difficult to capture in law. The GDPR contains the risk-based approach but what is still lacking a bit in the academic debate is the discussion of what is the impact of a privacy infringement/what is the damage of someone knowing some information: Sometimes it is easy to de-anonymize, sometimes it is hard work (e.g. the case of AOL search terms); Encryption can always be broken if you try enough. I would hope to quantify the risk to privacy a similar way: what information is safe in the future and up till what point. Maybe the safest approach at the

moment is to take the risk-based approach (what could go wrong; what information could be inferred; and how harmful is that). An additional challenge is that the GDPR assumes certain fines but it says little about quantifying the damage people suffer. The fines also don't take into account if you took some measures at least, of course the DPA's have some flexibility there but the law does not. We should balance the costs of measures that were taken to try and comply and the damage that was caused.

Question 6 –We have personal data and within that concept we also have sensitive data. In scholarly debate we have discussion on what sensitive data should include exactly. Perhaps we could apply the same to purposes, e.g. take a more processed based approach. How does that relate to anonymity in your opinion?

- The process is not linked to anonymity itself but the process determines the utility of data. So, for a certain purpose you need a certain utility, which requires a certain type of information to be contained in the data. Thus, if that information is not contained in those data, you cannot use those for that specific process. The same data could have utility for another purpose, such a purpose could be re-identification. We don't want to say that you cannot use certain data, but rather that they should not be used for a specific purpose. The problem is that you cannot distinguish between the use/purpose, once information is extracted it could be used for 'good or bad'. In information security we have homomorphic encryption: then you can process the data and only see the information that you need to see for that purpose. That is the ideal case, but requires a level of security. Another example is the use of secure multiparty computation. Sometimes that is feasible, sometimes not.

Question 7 – Could we say that a lot of these problems are caused by the GDPR focusing on the collection of data rather than regulating the use of data (while harm might come from the use of data rather than the collection)?

- True, but the problem is the use of course does not just harm, we also use data for many benefits. We want those benefits without the harm. So, we can use encryption for example to reduce harm, so that you only see the information that you need to see (we have attributed based encryption, attribute-based credentials, etc.) and shape it for a specific purpose, perhaps that is where we should be heading.

Question 8 - If we would from a legal perspective have a prohibition that data cannot be processed for a certain purpose, could we still use the notions then of anonymity or pseudonymity so that data can still be processed for a certain purpose because it is really necessary?

- The problem with the purposes would be is that data can still be used for good or bad and even if you have a law that says you cannot do it for that process people will still try to use it for another purpose. It is very challenging to enumerate the purposes and there are always new ways to process data in new purposes.

Question 9 – You have written about big data and cloud computing, do you think the changes in the past years in the data landscape (let's say between the DPD and GDPR) are something that we should take into account when regulating data, and what are those challenges? Can we still hold on to the assumptions on cloud computing or not so much?

- Something has changed since then certainly. The task that cloud companies provide is mostly linkage. Now we have centralization on the big cloud providers and that was not anticipated so much in the time of the 'early Googles'. This poses challenges, e.g. the concepts of data controllers is hard to maintain in a blockchain system. It is very hard for law to anticipate such developments, that is why the GDPR has to refer to the state of the art. But what is the state of the art: is it what the big players do or something else?

7.1.5 Kissim

Interview 19-04-2022
K. Nissim – Georgetown University (topic: differential privacy)

- Introductory thoughts/suggestions from the interviewee: There are some points to take into account for a legislative approach, from a point of view of computer science. It is important to take a pro-active approach to legislation. A reactive approach enhances vulnerabilities instead of providing protection. From a computer science point of view, one can say that once information is public there is no going back, therefore it is important to think pro-actively about putting information out. Maybe in that regard the world could be treated as more adversarial instead of just incentivizing to handle data well.

Question 1 – Can you introduce differential privacy to us? What does it do exactly, to put it in legal terms, for protecting the identity of individuals?

- Differential privacy captures only a sector of the concept of privacy. You could view differential privacy as a criterium that says something about whether an algorithm preserves privacy or not (using the term privacy in a broad way). Differential privacy allows for computations about individuals in the sense that the influence of the individual's data on the output is pre-determined. The outcome distribution is bounded, it is controlled. It is based on the assumption that, if somebody participates in the computation, the outcome of said computation will be the same whether that person contributed with his data or not, irrespective of the correctness or incorrectness of said data. With differential privacy you cannot learn more information about an individual that you would not have known if they did not bring their data into the process. Differential privacy allows for statistical computation or analysis while protecting the privacy of individuals, as the output does not show a specific individual's participation.
- Differential privacy is relatively speaking still quite a new field, so it can take some more years to see its limitations, accuracy and for example what level of noise is the most efficient, and thus for its applications to improve.
- One of the advantages of differential privacy is that it has a parameter that measures privacy laws. Every computation leaks more information, some data may itself not be so revealing but together with other data can be very revealing. Potentially to the point that you can start recovering information of specific individuals, reconstruction. Each use of data comes at a cost of privacy, we can do a form of an accounting process to determine how much privacy loss there is for individuals (privacy budget). Together, this forms a composition effect to the use of data. This is an aspect that is important to take into account in the regulation of data.

Question 2 – Speaking of the composition problem, is there a mathematical way to measure composability?

- That is difficult to say in such a general way. We have the parameter epsilon in the simplest calculus, you can add up the epsilons. It has to be bound as tightly as possible, so as not to exhaust the privacy budget very fast. With respect to other technologies, such as k-anonymity or other technologies that do not have a full mathematical framework such as with differential privacy, people have not looked at composition so closely. With respect to some of them, for example k-anonymity, we know that it does not compose. E.g. if you anonymize the same dataset twice, you have two k-anonymous tables, but when you look at them together you would lose k-anonymity all together. (More recently that theory has been proven in practice, see the work of A. Cohen & K. Nissim on this.)

Question 3 – Do you think having a mathematical definition of singling out, linkage attacks, or inference, could we already state that data would be anonymous just by complying with the notion of

singling out? Or differently put, do you think that if we would have three mechanisms to calculate the degree of anonymization, such as a threshold for singling out, linkage and information inference, this would provide not perfect but sufficient guarantees to assume that data has been effectively anonymized?

- It is not preferable to use the term anonymous/anonymity, as what anonymity is, is not perfectly defined. It has not even a partially mathematical definition. So, your question reflects on whether the notions of linkability, singling out and inference are going to protect us in the future. It depends on definitions, e.g. we could define protection against inference and we could adapt the definition of differential privacy. Is that a good idea? Perhaps not. Whatever definition we have for inference, differential privacy should fit that definition, because it offers strong protection against inference. The A29WP definitions are not technical enough. With respect to inference, there is chance that we can have a definition that is composable, which obviously would be good because it would allow us to account for the privacy loss. With respect to singling out, as we write in one of our papers, our attempt to formalize it will not result in a concept that composes. If we would still push for a definition that composes it could just lead to the unsatisfactory result of saying 'it should satisfy differential privacy and that is it. So, in our paper we aim to respect the opinion of the A29WP but give more meaning to it, but no matter who defines the notion of singling out (whether it is our own definition or that of the A29WP) it will probably not compose.

Question 4 - in your paper towards formalizing the GDPRs notion of single out you assume that the predicate singling out refers to the original dataset, not the released dataset, i.e. it is the ability to single out a row in the original dataset from the information of the released dataset. Why did you not consider singling out in the released dataset?

- The data that you want to protect is the original data, not the data that is being published. Look at the underlying data, that needs to be protected, that is a point that was missed by the A29WP. The result of this stance of the A29WP is for instance that in the analysis they considered k-anonymity as protective, because if you just look at the outcome of k-anonymity and you forget how that outcome was produced, it may look safe. But if you do not forget how the outcome was produced, you are not safe. The work of A. Cohen demonstrates that in a very strong way for example, for a large group of k-anonymous mechanisms you can reconstruct some of the original data just by looking at the outcome. So, there are two points to make here: the first is that what you want to protect is the underlying data. The analysis output, you may or may not want to protect. The second thing is that when you are talking data protection, you need to take into account the process that takes the original data of individuals and creates the release. Because if you know the transformation you can deduce the input. Because of that, differential privacy looks not at the outcome of the analysis but rather at the analysis itself, whether the analysis satisfies the differential privacy criteria. This criterion limits the informational relationships between the input and output to protect information that is specific to individuals.

Question 5 – Do you think we should extend the notion of singling out in the EU? Within the context of the GDPR it is about being able to isolate a record within a group of records. It seems that your approach to singling out is more about 'guessing' if a record can be in the original dataset. Should we keep a narrow concept of singling out or should we see it in a broader way?

- Obviously, there is a reason why the notion of singling out was included in the 'holy trinity' (singling out, linkability and inference). It is a useful privacy breach to present to people, so they are aware that these are the type of privacy 'failures' that we want to prevent. I would not choose singling out as a criterium, but I respect that decision of the regulator. What you are describing in your question refers to membership attacks. Differential privacy protects against those attacks. Membership attacks can be protected against in a good enough definition of inference. Membership attacks are more related to inference than to singling out. Again, here we come to the point of the A29WP focusing on singling out in the output, which I believe is not the right

approach. It seems if we want to protect individuals, we should protect the information that they provide, if we protect the output what are we protecting? We should apply the principles to the information that we want to protect, rather than focusing on something of which we generally do not even know how it was created (which is the outcome or release of the analysis. You also mention isolation. Isolation in itself is an insufficient criterium, because you can 'isolate' without having access to any data, it is just at random and you can succeed with high probability. We need to supplement it with another criteria.

Question 6 – In your paper 'Data Protection Composition Problem', you find that the Census Bureau underestimated the re-identification risk by a factor of 4500 in light of the initial identification rate of 0.0038%. In your view, how should a data controller assess the re-identification risk? Or more simply put, what should be sufficient for a controller before they release their dataset?

- So, we understand we need sufficient measures (e.g. differential privacy), but it does not tell us what is sufficient or what the controller has to do in that regard. There is a not a clean cut answer to the question of what a data controller should do to give these insurances. If controllers use specific technologies they get some assurance from the legal framework, but you cannot always use those technologies in every situation as it limits the use of the data. What is worrisome are instances where controllers can use loopholes. For example, claiming to process data for statistical purposes to be guaranteed (partial)exceptions under the GDPR. The regulator needs to add constraints and responsibilities on controllers for controllers to show why it is safe to use for example statistical data processing, some kind of analysis and insurance that scenarios are taken into account related to releases for example. How to incentivize the controllers to think about this? We need to be careful with declaring some releases as structurally being safe and bring composition awareness to the controllers. We need a justification that can stand scrutiny before information is released. The GDPR at least has the most advanced approach (e.g. compared to the USA) in the sense that it has definitions and it has meat that we can analyze and try to improve. But it also has its loopholes, such as statistical processing. It is important to look at those loopholes, such as anonymization and justifications such as research purposes and assess whether those are really needed in the situation. So, we need to somehow create incentives to do that in the right way.

Question 7: As previously mentioned, we have this tension between heuristic methods and formal definitions of privacy. The GDPR seems to be more focused on the formal, semantic, approaches of this problem, and it does not specify any types of methods. This can have legal uncertainty as a drawback because you need to interpret concepts. Should we transfer to process-based approaches or stick to the current approach? From your discipline/perspective, what would be preferable do you think?

- When we can use formal definitions (of privacy) this has a lot of benefits, we have a chance to understand what we are doing. With more ad hoc heuristic methods we can never have this level of understanding. It is good that the GDPR does not mention for example different techniques such as differential privacy, so that we do not hardwire techniques that later turn out to be faulty. Differential privacy is quite future proof, but one should not hardwire it. There could be other better ways to go about this, for example with best practices, or practices that could be incentivized when possible, by regulatory bodies. Perhaps updated guidance and more frequently having updates from the EDPB would also be helpful.

7.1.6 Lefkovitz

Interview 10-05-2022
N. Lefkovitz - Senior Privacy Policy Advisor NIST

Question 1 – In current society with large volumes of data available and techniques to extract information from those data, how do you view maintaining anonymity or privacy protection? Is indeed easier to identify individuals? Do you see that challenge from your perspective?

- Absolutely, we can see a mosaic effect. We aim to bring a metrology approach to data processing and privacy. We ultimately at NIST came up with complimentary privacy engineering objectives in addition to the concepts of confidentiality, integrity and availability, to also offer more tools to engineers. We have three privacy engineering objectives you can find them in our NIST privacy framework for example, but perhaps the most important of the three for this discussion is the objective of disassociability. The objective is to disassociate information from identity and devices beyond operational requirements. At least as much as possible. Similar to security objectives, there is always a degree, and you are going to trade-off with other objectives as well. But it gives engineers a sense of what the objectives are so that they can build capabilities. In addition to the objectives, we also developed a complementary risk model. This allows us to have discussions such as we have in the domain of security, e.g. how to mitigate risks, how to trade-off costs and performance, etc. Because we see privacy engineering and risk management on a spectrum, while anonymity as term of metrology is okay, we know what a perfect state is, but we don't expect to get there. You can never mitigate a risk to zero. No perfect security, no perfect privacy. It is a balance that we have to find among organizational goals, individual needs, regulators and societal interests and so on. Anonymity in a legal sense is not very helpful, it makes it hard to do that spectrum of reasonable measures. That is why we choose disassociability, then you can think of which level of identifiability is acceptable and necessary, but still getting the functionality.

Question 2 – Can elaborate a bit more on the concept of disassociability that you describe in your Privacy Framework?

- We wanted to add other constructs in addition to the Fair Information Principles. One example is that we have a category named disassociating processing. But is important to remember that in this document we focus on outcomes, not so much about specific techniques/technologies. In cybersecurity we have more maturity of concepts, so for the privacy framework we use more examples to explain what we mean and what techniques you might use to achieve a certain outcome, so those are meant to explain but they are not meant as a limitation. We capture the limit of identifiability we are not trying to set a binary bar; it is a spectrum to think about different outcomes. You might use different techniques, differential privacy, tokenization, privacy preserving crypto, etc., and you can outline them as well. How you might achieve disassociability has to do with data protection solutions which increase disassociability consistent with the organization's risk strategy and enable implementation of privacy principles like data minimization. And we wanted to recognize that there is a view that data minimization more than just a principle of limiting information, but actually encompasses a whole range of techniques rather than just minimal data.

Question 3 – There are so many privacy preserving techniques out there and of course you also mention that sometimes you can use a combination of techniques. Yet, at least in Europe, we do see some debate that perhaps some techniques are not strong enough anymore in this day and age and that there are some techniques that are more robust. Could you say some techniques are stronger than other techniques in general or is it really context dependent?

- I would agree more with the latter. Some techniques give stronger privacy properties than others but at the same time can come at a higher cost in terms of expertise and performance, to the

extent that for example k-anonymity which is not as robust as differential privacy, is perfectly useful in some cases. We should not regulate specific technologies but focus on the outcomes and allow organizations to select the technologies that work towards that outcome. Regulators can assess whether that is sufficient enough or not. But if one would regulate technologies, it is either now or in the future not going to work. It is also a question of implementation, you can have all of these techniques but have a poor implementation. So, it is important to regulate the outcome and focus on proper implementation. We are also paying more attention to research and applied solutions so we can put out guidelines and so we can build standards and certification. That will help with maturing the technologies and getting more widespread adoption.

Question 4 – Under the GDPR we also have the concept of pseudonymous data where you have personal data about a person, but you remove direct identifiers to offer a form of technical protection, so that you would need some additional information to identify that person. How do you view the concept of pseudonymization or pseudonymized data?

- This relates to risk management. There was, in contrast to cybersecurity, no consistent risk model and thus no tools for organizations to assess privacy risks. In the privacy framework we have a venndiagram on how cybersecurity can manage privacy risks. Organizations need to process data and as a side effect of that processing there can be privacy risks (e.g. harms, problems, it is a broad range). Threat is a meta value in a way. We have the idea of problematic data actions: operations with data that create some sort of problem or harm for individuals, e.g. dignity loss, stigmatization, discrimination, loss of self-autonomy. With this meta factor you can ask the question whether this collection of data is likely to cause any harm and what the impact would be. You don't have to classify data in any category, you don't have to get into definitions of what is personal data and what is not for example. Instead, you say what you are doing and what are the contextual actors and likely outcomes. That model and tool enables a discussion and makes it concrete. So, once you understand the risk you can decide what the mitigations should be. Rather than creating categories, think of pseudonymization as a technique and decide based on the risk whether it is a suitable measure or not.

Question 5 – You mentioned contextual factors that determine the risk. Would you in those contextual factors also include the type of actor? (governmental actors, private sector bigger and smaller actors, e.g.). Not every actor has the same means in terms of know-how, budget, etc. to protect data.

- Maybe we should think more along the lines of if you do not have the means to protect, you should not be undertaking that activity. That is why you for example have smaller actors that go to vendors who offer security protection. Also not even all small businesses are equal, context is even relevant there.

Follow up question: the question also concerns who are you applying the norm to, e.g. some actors might be able to identify an individual in a dataset while another actor is not able to. So, who are we addressing the norm to?

- If the organization cannot identify them in the processing activity it might be reasonable, but if are disclosing those data or putting them in some sort of data pool, or if the data goes into training models/is shared for training model, maybe it is not so reasonable anymore. So our risk model is intended to capture those aspects, e.g. an action might be fine now but actors are also asked to think of the likelihood of re-identification.

Question 6 – Looking towards the future. Is NIST also exploring the possibilities of synthetic data?

- Yes, we have done challenges on those, for example within differential privacy challenges. With differential privacy algorithms you can come out with strong sets of synthetic data. It is just a technique in the toolbox. E.g. federated learning combined with privacy enhancing cryptography

such as homomorphic encryption, could allow models to be trained on raw data in a well-protected manner.

Question 7 – Where are we now in terms of the state of the art? What are the challenges now in terms of protecting against being identified?

- In 2019 we launched the privacy engineering collaboration space, which is open to the public. We started from de-identification and that turned into differential privacy. The tools for that are a little bit more mature, but then we had the tools but not yet a lot of activity. So, we started a blog to inform people on what differential privacy can be used for and so on. Now we hope to turn that blog series into technical guidelines. And of course, we are not the only ones looking into differential privacy. But differential privacy is probably leading. Often, we put out guidelines we can take those further to standardization organizations and from standards you can get to third party certification. Then vendors can make more robust assertions about what their products do or don't do. And of course, differential privacy is not the only tool. So, we see the path that we would take but we are still at the beginning of that path. Activities such as challenges and pilots also help in learning and set us on that path to standards and so forth.

Question 8 – How do you incentivize actors that handle data to apply protective measures?

- A risk based regulatory approach helps. You have to implement reasonable measures that align with risk. There is a level of uncertainty, you cannot game definitions in a way, you constantly have to take reasonable measures. There might be tools out there that are suddenly reasonable in terms of cost. We might need a layered approach in rulemaking: for example more regulated types of access or areas or having sunset provisions. But a foundation of risk management is probably more efficient than prescribing specific requirements.

7.1.7 Limniotis

Interview 19-04-2022 K. Limniotis* - Department of Informatics and Telecommunications, National and Kapodistrian University of Athens/Hellenic Data Protection Authority (topic: pseudonymization and encryption techniques)

Question 1 – Before we delve into the technology, I have one question about the concept or definition: in your webinar in December, you briefly mentioned that there is a definition of what pseudonymization commonly means among engineers before the GDPR and then the GDPR definition of pseudonymization. Is there any difference between what an engineer would see as pseudonymization in a technical sense and what the GDPR describes/requires?

- Both the definitions or concepts (the technical and legal aspect) are very close. From the technical perspective it is commonly agreed that pseudonymization means replacing one or more identifiers with a pseudonym. This is in line with the legal definition of the GDPR. However, interestingly enough, the definition in the GDPR does not refer to the word pseudonym. Pseudonymization is defined there as processing where additional information, that allow re-identification, are stored in a different place; thus we may have pseudonymous data without having an explicit pseudonym. This can be, e.g., illustrated by the famous example of the hospital in the USA years ago (see Sweeney's work[289]) where data such as age, diagnosis, address data etc. of patients were available online, because they were assumed to be anonymous. However, this turned out to be not the case, as with other public information the researcher managed to re-identify many of them. In this example there was no use of explicit pseudonyms. But at the same time these data could be considered as pseudonymous data under the GDPR, as they are certainly not anonymous data. In that light, the two definitions are very close, but not fully identical.

Question 2- Does the same go for the concept of anonymous data? Are the GDPR and technical concept of anonymous data very similar?

- For anonymous data the concepts are more similar. There is a difference though. Many engineers assume levels of anonymity, for example partial or full anonymity. The GDPR is stricter on this; either we have anonymous data or not. If it is difficult to identify individuals but not impossible, we should not refer to those data as anonymous data, from a legal point of view. However, for example an engineer could say that we have almost fully anonymous data.

Question 3- In your opinion what would be the better approach: the one that the GDPR takes, which is more a black or white stance, or the more technical approach where you could see anonymity as more of a sliding scale?

- I would prefer the notion of the GDPR. That is why pseudonymization becomes even more important. Under the framework of the GDPR it is very difficult to achieve anonymization; not impossible but very difficult indeed. If we have though a nice pseudonymization procedure, data protection risks are diminished. Thus, the stricter concept of anonymization is somehow to some extent balanced by having pseudonymization. The notion of pseudonymization is expressed multiple times in the GDPR as an appropriate safeguard for data minimization, security reasons and so on. Hence, even if a data controller cannot ensure anonymization due to the inherent difficulty of this task, data protection risks can be eliminated by a robust pseudonymization.

Question 4 – An interesting category of data is that of statistical data, which can for example be very important in the public interest, for example in policy making. In your webinar you briefly mentioned

*The statements in this interview reflect personal views and they should not be considered as statements that reflect the views of the Hellenic DPA.

[289] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

that pseudonymization is popular to use on statistical or aggregated data. Can you explain a bit more about that?

- Think of the example in which the actor that has to conduct the statistical analysis needs to have access to the original data in order to perform the analysis. In this example, if a researcher or university wants to conduct an analysis, they have to ask for the original data from the public entity that has that data. If that entity would provide the full data to the researcher, that would not be in line with the data minimization principle, as the researcher does not need to know the exact identities of the data subjects. Here pseudonymization is useful, so as to ensure that statistical/scientific analysis will be conducted on pseudonymized data. Of course, even in this scenario when those data are provided, we cannot fully exclude that people within that data will still be identified, for example with other data that are available out there, thus such data would not be anonymous. That's why our aim should be to have a robust pseudonymization.

Question 5 – If we can distinguish between the more classic and advanced means of cryptography. We encountered a lot of criticism so far on more basic methods of pseudonymization, such as hashing, as not being enough to achieve pseudonymization under the GDPR without a combination of other techniques or simply because there are identification risks. Do you think not very advanced techniques such as hashing are adequate means of pseudonymization to comply with the GDPR or not so much?

- Very simply put, I would agree with that criticism, but it is more complicated than that. For privacy enhancing technologies we should keep in mind the GDPR adopts a risk-based approach. For hashing for example, one cannot exclude a scenario in which it can be useful, as there could be a scenario with very low data protection risks. However, indeed the use of classic hashing (i.e. where we do not have a secret key) for pseudonymization has significant drawbacks, as one having some knowledge of the pool of input may quite easily re-identify some people from the pseudonymized dataset. But as I said, we cannot exclude hashing altogether – and actually hashing with a secret key alleviates this issue. In any case, the GDPR requires the controller to prove that the necessary data protection guarantees are in place, related to the relevant risks.

Question 6 – Staying on the topic of the more classic means of cryptography. Are there other forms of more classic, not so advanced, cryptography, that are perhaps stronger than hashing?

- To clarify, the presentation in the webinar referred to before was focused on cryptography for pseudonymization, but of course there are also other means of pseudonymization that are not cryptography-based techniques. If we stay on the topic of cryptography, then indeed if we use a secure robust cryptographic algorithm, risks of identification in terms of reversing the pseudonym back to the original identifier are very small. This is because these risks are related to the risk of breaking the cipher, hence that is why it is so important to have a robust cipher. Of course, we should take into account the state of the art, as something that is secure now might not be so in, for example, 10 years or even earlier. And either when we use classical symmetric or asymmetric cryptography, we can have practical issues, such as huge pseudonyms that are not easily comprehensible by people. But in essence, classical cryptography is good from a security point of view, yes. In such a case the additional information that allows the reidentification according to the GDPR, is the encryption key, thus it is important to protect that key.

Question 7 - Do you think it would be useful if we have a legal regime that distinguishes between a third-party gaining access to the key through an attack for example and breaking pseudonymization that way or on the other hand internal re-identification? So, is either external or internal re-identification a bigger risk and should and instrument like the GDPR distinguish between the two?

- The GDPR indirectly covers both scenarios. If the organization for legitimate purposes needs access to the direct identifiers and if the organization needs also to provide pseudonymized data to any other organization, this would fall under the first scenario. But the GDPR also indirectly

covers the second scenario. An organization usually has multiple departments, and only some of those departments need to have access to the original identifiers of personal data that the organization processed. It is a matter of security measures to ensure that internally not every person has access to all data – actually, each person should have access to personal data on the basis of the 'need-to-know' principle, depending on her role within the organization. Security measures under the GDPR do not refer only to external adversaries, but also to internal adversaries. For example, when DPAs have to make an assessment that also includes scenarios of internal adversaries.

Question 8 – Does the GDPR offer enough incentives to implement pseudonymization or other protective measures? For example, because if a controller implements pseudonymization they more easily comply with their obligations under the GDPR.

- Of course, the GDPR is about a fundamental human right. So ideally, we should not expect incentives to respect fundamental rights. In that sense, I think that the GDPR could not provide better incentives or an alternative. In my opinion, it is probably not be the role of legislation to create incentives. However, in this direction, we should recall that some obligations of the data controllers are somehow, roughly speaking, 'relaxed' by implementing pseudonymization or other protective measures; for example, is such safeguards are in place, the controller may not have an obligation to notify the competent DPA in case of a data breach that does not entail risks for the affected persons. Moreover, the GDPR promotes accountability tools, such as codes of conduct and certifications. Although they are not obligatory, the GDPR encourages the adoption of such tools which need to be 'approved' by DPAs. Pseudonymization or other protection measures may be part of these accountability tools and, thus, if pseudonymization is to be implemented, it will be checked if it is strong. The GDPR also refers to the positive impact of adhering to or applying certification and codes of conduct.

Follow up question: should the GDPR distinguish between doing the minimum that is necessary in a certain situation to comply with pseudonymization requirements and between offering stronger protection, which is more difficult for the controller?

- The crucial point is that the GDPR follows a risk-based approach. This means that it is very difficult to speak of minimum technical requirements. What is not enough protection for one case might be too excessive for another case, depending on the relevant risks for each case. We should also have in our mind that. some types of data processing are not possible with pseudonymous data, as well as that each time we need to take into account the state of the art. Relevant guidelines are for sure very important, but the legal instrument itself should not enforce minimum technical standards. This also allows some flexibility in the future, for example also taking into account quantum computing.

Question 9 – As an example of more advanced techniques you mention user generated pseudonyms. Can you tell us a more about that: in which situations or contexts are those most useful and what are some limitations in using them?

- In some scenarios there is no need for an organization to have direct access to original identifiers of the users whose data are being processed, in accordance with the data minimization principle. In such a scenario the user could generate her own pseudonym. In these cases, one crucial issue is, how does the user prove that she/he is the owner of the actual pseudonym, if she/he wants to do so. If the user is able to do this, the additional information required for identification under the definition of GDPR is in the hands of the user and this would comply with data minimization. To give an example: There is a Greek electronic ticket system for public transportations that monitors the 'travel of tickets' for well determined legitimate purposes, but not the travel of passengers in a personalized way because the competent organization does not need to know that specific information. Hence, in this scenario, the organization 'monitors' only the movements of tickets; each ticket has a specific number and the organization does not know the

identity of the ticket holder. However, this number (playing actually the role of a pseudonym) is cryptographically generated by the passenger herself (via a process determined by the public transportation organization). If a passenger wants to prove that she owns a specific ticket (e.g. if she has lost it), then she can prove to the organization that she is the owner of the ticket with the specific number. Hence, data minimization is fulfilled in this case by user-generated pseudonyms._. In general, there are several challenges in designing such a system. For example, we need to ensure that two users do not generate the same pseudonym, and we need to ensure that nobody can prove that owns a pseudonym that she does not actually own.

Question 10 - Which means of cryptography are strong techniques that can protect against de-pseudonymization for some time to come that we should look into as good practices?

- In terms of classical cryptography, symmetric encryption is still a nice option. It looks like it will also be still strong in the future (e.g. in 10 years) and is not affected by quantum computing, as long as the key size is appropriate. Asymmetric encryption -which is mostly known as public key cryptography-, includes ciphers that are widely used right now as secure, but they are not post-quantum secure (i.e. they will no provide security against attacks by a quantum computer). On the other side, many processes need the properties of asymmetric encryption. So, we need to focus on ciphers that are already post-quantum resistant. But of course, the question is for what time period we want to ensure security: is it for only a couple of years or ten years? (for example).
- Follow up question: in previous research, techniques that surfaced that we were advised to focus on where techniques such as zero knowledge proofs and homomorphic encryption. Are such techniques strong forms of protection in your opinion?
    - Absolutely. Zero knowledge proofs, secure multiparty computation/secret sharing, homomorphic encryption etc. should be definitely taken into account when evaluating the proper security/pseudonymization methods. It is important to know that such techniques have been known for years in the cryptographic community; they are also known with the term 'privacy enhancing cryptography.' But they have not been considered or described much as tools, e.g., for pseudonymization in relation with the GDPR requirements. If we look at the protection they provide and the description of pseudonymization under the GDPR, these advanced cryptographic techniques are actually fully compliant.

Question 11 - Do you think it still makes sense to distinguish between different types of data along the aspect of identifiability as the GDPR currently does? Thus, having anonymous data not in the GDPR, and in the GDPR differentiation between pseudonymous and non-pseudonymous personal data and stricter requirements for sensitive data.

- Sensitive data is essential, this is a good distinction from a human rights perspective. The crucial point is the definition of anonymous data, this is the difficult issue. How can we be fully sure that we are dealing with anonymous data? This could be further clarified. The GDPR already does its best of describing it (see the recitals on anonymity); I am not sure how to make this even more detailed in a legal instrument. Perhaps we could further elaborate on other means such as guidelines or technical standards for organizations to address in practice the notions of anonymous data, anonymization and the relevant data protection risks.

7.1.8 Sandbrink

Interview 25-04-2022
K. Sandbrink – National Cyber Security Center

Question 1 - In your field of work, what are technologies that offer a good privacy protection (in the sense of protecting identity, but it can also be broader than that)?

- Most of the technology I work with are enhancing cybersecurity in general and of course data protection forms a part of cybersecurity. So those technologies are not specifically targeted at data protection, but they all contribute to it. Almost any measures or safeguards such as encryption, strong forms of authentication, strong access control, vulnerability management, etc., all those contribute in a sense to data protection.

Follow up question: To be more specific if we are looking at specific techniques, such as encryption methods (secret sharing, homomorphic encryption etc.), which techniques should we focus on most for strong identity protection?

- An important aspect no matter what technique or technology is used the standardization of techniques and examining new proposed techniques. For example, a new form of encryption is in practice perhaps not the best encryption. Thus, it is most safe to focus on encryption algorithms that have already proven themselves and are standardized, such as AES or RSA algorithms.

Question 2 - So a step 0 is what are good techniques and in that sense standardization is an important process. Do you see any challenges in finding a technique that can be standardized more so now than a couple of years ago?

- Not necessarily right now, encryption is always a race against the clock in cybersecurity where technological possibilities increase so cybersecurity has to improve as well. One of the challenges in terms of technology is quantum computing, as data still need to be adequately protected in 20 years' time. We have to encrypt our data having this in mind, but therein also lies a challenge as it is not exactly clear yet what will be possible with quantum computing in the long term. So right now everything can be protected relatively well but in the near future that might change.

Question 3 - Given that challenge, what is important to take into account when preparing for quantum computing?

- To achieve the best protection, we need to use state of the art encryption, so again, RSA/AES algorithms and make sure the key lengths are sufficient. For quantum computing you need more bits than the current standard, so those are heavier applications of the algorithms than the current applications of algorithms. Also keeping up with the more theoretical research not just applied research, on what quantum computers might be capable of and what types of algorithms might be the most vulnerable to be cracked. So for example, elliptic curve algorithms seem to be the most resistant to quantum computing, (discrete logarithm algorithms, less so for example).

Follow up question: The research that you mention, does that take place on the international level?

- Yes, it is an international debate and also an arms race between countries in developing quantum computing.

Question 4 – You also advice on privacy risks. In your view & terminology, what can we understand as privacy risks?

- The first thing is always data minimization, thus the question of whether you really need all the data that you are gathering, that is always the first step. Of course, that is also in the GDPR, but

it is often overlooked. Usually there is more emphasis on how to protect or share data. The data minimization part gets skipped often, the question of whether you really need all that data. And of course, advising on the protection of data, you need all the cybersecurity measures that are needed.

Question 5 – Nowadays there can be a lot of strain on the principle of data minimization (in this datafied/data-driven society) and there is a push for more data collection. To what extent do you think maintaining data minimization is still doable?

- Often in the context of big data analytics and the like, organizations gather data before they know what to do with it. From a data science perspective this makes sense because you want to gather as much data as possible to analyze and research. It does not mean that you should just gather all the data because there might be possible innovation. You still need to think about the purpose first, so why you want or need certain data.

Question 6 – Along that line of thinking, is it difficult to keep data separated? Or is that not really a challenge in your field of expertise?

- It is difficult to keep data sperate. Because you can make profiles out of combined data it is more important to keep data separate and to limit data sharing between organizations.

Follow up question: So, do you think there should be stricter rules on data sharing to prevent compilation of data or would that not be a good idea?

- More restrictions on data would be a good idea. The current restrictions are a bit hard to work with if you just have an IT supplier and they need to process data and you need specific agreements. It can be complicated for some organizations that they need to make those agreements again and again. This can make it difficult for individual people in organizations and they might try to share personal data with other organizations and try to dodge the rules a bit. Thus, there should be proper restrictions, but it should also remain possible when it is justified to share data with reasonable measures without overburdening organizations or incentivizing them to avoid applying the rules.

Question 7 – Do the concepts of anonymized data vs. non-anonymized data, and personal data that are protected with pseudonymization vs. personal data that are not protected in that way, play a role in your work? Or do you define categories of data very differently?

- It is not very relevant in a way; we also work with non-personal data. We do have guidelines on securing web applications, they do mention them, but they are not the focal part of our work.

Question 8 – If we would not use the concepts of having anonymous or non-anonymous data and having personal pseudonymized and non-pseudonymized personal data, what would be important concepts in your work? Thus, if we would not focus the protection so much on the identifiability aspect, but rather on other aspects, what would be useful aspects in your work?

- We view data as having two states. Data can either be at rest (stored) or data can be in transport (e.g. being shared over the internet). The distinction is relevant to what the risks at the moment are. A lot of protection of data in cyber security focusses on data in transit, because it is a more imaginable risk that data get intercepted. However, data is at rest most of its lifetime, almost 99% of the time. It takes some effort to imagine where there is the most risk, if it is in transit for example. And it is not a sharp point of distinction, where do data go from being at rest to being in transit? If you want to send it for example there are multiple moments in the process where goes from being at rest to being in transit and vice versa. The extremities are clear, but the in-between situations are less clear.

Follow up question: you mention that most of the protection is focused on data in transit, while data spends most of its time at rest. Do you think there should be more measures or more focus for data at rest?

- Encryption of data at rest is getting more attention. But the focus is on the hardware not always on the encryption of servers. Thus, if there is a data breach at a cloud provider for example, all the data is plain text readable. Storage encryption in general should be the norm at least.

Question 9 – In your work do you focus mainly on the malicious side of risks (such as attacks) or also on privacy risks that arise in another way?

- Indeed, we focus on the malicious side of privacy risk, of course the other problems are in the scope of the DPA. We focus on three aspects of data: confidentiality, integrity, availability of data. Privacy is a confidentiality issue, the risk to confidentiality is usually malicious. And of course, there could always be human mistakes, but those are not necessarily a cyber security issue.

Question 10 – Are there any other future challenges in the coming years, next to quantum computing, that have an impact on how you might protect data at the NCSC?

- Quantum computing is probably the only technological risk. From a societal or political perspective, there can be risks such as governments weakening protection for law enforcement purposes. Or a lack of awareness among users about their own privacy. Users are often giving their permission for many aspects and to many actors without really understanding what they are giving permission for. This may lead to companies harvesting so much information, that by the time it is clear that there is an issue they already have all of this information. So it is important to increase awareness.

7.1.9 Sangers

Interview 13-04-2022
A. Sangers – TNO (multi-party computation)

Question 1 - What really is multi-party computation (MPC) and how does it protect the identity of individuals in the data? Is it more of an anonymization or pseudonymization tool?

- You can see MPC as a functionality. It can be used on different data sources from different parties without the other organization seeing anything else than the output. The protocol or technology used for the MPC can differ per situation, it can be used with or without a third trusted party. MPC is cryptographically/mathematically enforced, it is less dependent on organizational measures. Most commonly used in MPC is secret sharing. Within MPC you also have different models or levels of strictness possible (in that aspect one can compare it to blockchain). For example, active security is the strictest form. Thus, from a technical point of view through MPC you can offer very strong protection.
- MPC does not offer any privacy guarantees for the output of the processing, the output can contain personal data. MPC protects input and intermediate results.
- There can be a different number of parties providing input and individuals represented in the data. In terms of scale, a couple of input parties are already enough to use MPC.
- The distinction between anonymous or pseudonymous data is more a legal distinction. There is not a zero or one answer to the question if the data are pseudonymous or anonymous. While the GDPR represents a black or white regime, anonymity from a technical point of view is a sliding scale, as the technical measures can be applied offer a scale to data being more anonymous/pseudonymous or not. For example, in the context of differential privacy there is a choice in how much noise to apply. This applies mostly to the output, for the output there is really a sliding scale as to how to view the data in terms of privacy. For the input or throughput, we have a slightly different discussion. There are two approaches for the input/throughput level: the absolute or relative interpretation of what is identifiable. So, we have the three concepts of linkability, deducibility and singling out. Under the absolute approach of identifiability, true anonymization is never really possible. Under the relative approach which focuses on what is reasonably likely, we could achieve some degree of anonymity.

Question 2 - The Article 29 Working Party asserts that one should also prevent linkage and information inference. This is a topic you discuss in one of your papers. How does MPC prevent singling out, linkage or inference?

- In the paper that you refer to we took the example of synthetic health data, more specifically the scenario of a hospital and insurance company who combine their data to assess which patients have a high risk of heart failure. Firstly, you have to find patterns between the two data sources to develop a model. Second, you have to apply the model to new patients to assess risk. To conduct this you first have to combine the data somehow before you can look for patterns. We did this by using a third party in our example. We then used a combination of techniques including homomorphic encryption and secret sharing. Thus, how does one prevent an attack: One can use secret sharing and homomorphic encryption, those are the strongest forms of encryption. You can also use combinations of techniques to achieve the optimal result. With homomorphic encryption the data stay encrypted during the processing.

Question 3 -You mention a lot of different techniques and combining of techniques. To what extent can we see all of these techniques as MPC? And, in your opinion, to what extent can MPC itself be used to achieve anonymization?

- Essentially you could describe MPC as a toolbox of different possible technologies. Most people use MPC to refer to secret sharing or sometimes garbled circuits, for those techniques it has been

proven that theoretically all computations are possible in an MPC fashion. From a practical perspective sometimes it is better to use other technologies or a combination thereof, then you get more of a Privacy by Design approach, thus sometimes multiple PETs are used to get the same privacy guarantees. MPC is an established approach (it has already been around since the 80s), so theoretically a lot of computations are possible in an MPC way.

Question 4 -Could you say that MPC is essentially the model that allows us to share data inputs? And then for example homomorphic encryption is the model that allows us to further compute or distribute that data? Or are both aspects included in the concept of MPC?

- The confusion in terminology is because MPC is both a methodology and a technology. But it can be explained using a specific technique, such as secret sharing. Secret sharing is an alternative for homomorphic encryption, you could use one or the other. Sometimes you can use both techniques, but you cannot apply them at the same time. E.g. first use homomorphic encryption to combine the data in a secure way, and then for the machine learning model you could use secret sharing.

Question 5 - MPC is important for the input or intermediate output, but what would be a good way to protect the output of that process? Especially when it is not necessary for the purpose of the process to identify individuals in the output, because you are more focused on patterns, group profiles, etc., how could you protect that output?

- For the output phase differential privacy can be used to add noise. MPC has no utility drawbacks, there is no comprise on the utility to protect privacy, while it is sometimes argued for differential privacy that there is some balancing between the utility of the processing and the protection of privacy (the more noise one adds the less accurate the results will be). For MPC it is important to think about what it is in terms of an outcome that you need. For example, the model itself can be the goal. Or, in another example, the outcome can be to determine whether a patient is at high risk or not, for which you can keep the model secret, but you need to reveal the outcome of applying that model for an individual patient. Thus, the outcome or purpose determines what information has to be revealed, in principle you can apply MPC again and again until you arrive at the information that you need to reveal. Not all techniques are as privacy preserving as others and you can use a combination of techniques. For example, federated learning gives no privacy guarantees, but is better than combining all data, and you can combine it with homomorphic encryption.

Question 6 - Are there any limitations to MPC?

- It is not possible yet to do the most advanced models, at least the training phase. Some still take too much time for practical application.

Question 7 -Assumptions behind MPC are sometimes comparable to Blockchain, in that light: do you think that depending on the settings, one could categorize MPC as pseudonymization or as an anonymization technique? For instance, in passive setting, such as your paper, it seems that as long as $t < n/2$ you can achieve perfect security guarantees and output delivery. However, where $t < n$, only computational security but full output guarantees are possible. Do you think that this could have a say in how to categorize MPC, as a pseudonymization or anonymization technique?

- It is useful to have the classification of pseudonymization and anonymization. But in reality, it is more a scale. You can weaken or fortify protection, there is the number of adversaries, you invoke computational assumptions or not, you can choose for covert security as in-between active and passive security. Thus, there are a lot of options. If you really use the most strict approach then you could anonymize data. But of course it depends on the absolute or relative approach to identification. The absolute approach is not really usable, then you can hardly do anything useful with the data. The relative approach is more reasonable.

Question 8 - In which sectors or type of context is the use of MPC most useful or describable?

- It is especially useful to apply in the context of healthcare and finances (fraud, money laundering, 'know your customer' applications. Other government sectors and the energy sector are interesting sectors to use MPC in as well. Or more generally speaking for all slightly bigger organizations that deal with data it can be useful to make use of MPC. Also not just in protecting personal data, but also protecting confidentially.

Question 9 - Legally speaking we have a distinction between anonymous and pseudonymous data, but from a more technical perspective there is a sliding scale to identifiability depending on many factors. But which factors can we see on this scale, so which factors or aspects are important to take into account in determining identifiability?

- For personal data it makes sense to use the three criteria of identifiability. One could use MPC as a way to have anonymous data so that the GDPR does not apply once the data are anonymous. But in using MPC you already use a lot of protective measures, such as data minimization, decentral processing, adhering to the purpose limitation, etc. So, in a lot of ways using MPC is similar to applying GDPR standards and we can see the GDPR as offering very good guidelines. Thus, applying MPC as an anonymization technique so as not to have to comply with the GDPR does not make much sense. So rather it is not so much the distinction as in are data anonymous or not, but rather how can we protect them and in doing that the GDPR offers good guidelines.

Question 10 - Do you think the GDPR offers incentives in that sense to protect data (for example, using MPC already offers a lot of protection from a point of view of data protection by design)?

- Yes, and the question is rather how long it will be until the GDPR (or similar instruments) will require the implementation of PETs. In an indirect way one already has to use PETs to be compliant with a lot of the privacy requirements.

Question 11 - MPC requires some investments, does the GDPR offer enough incentive or reward for that protection?

- There should be incentives to improve the privacy of your processing. If the choice is only between anonymized or pseudonymized that is a hard discussion. On the other hand, in applying MPC you can sometimes do more with the data, because you are minimizing to the data that you need and the proportionality is improved etc. So, using MPC helps, but we should not view it in terms of being 'rewarded' with anonymization. With anonymous data, one is allowed to do much more with the data, for example the purpose limitation does not apply anymore/offer protection anymore.
- Speaking of the term anonymity itself. It is more relevant in the legal discourse, some technical experts are less reluctant to use the term anonymous/anonymity. From a technical view, data could be called anonymous data when just some variables are removed, so different than the legal/GDPR definition.

Question 12 - What are some other techniques that we should consider besides MPC?

- It would be good to have a list with other PET's that can be seen in addition to MPC. For example, homomorphic encryption, federated learning, trusted execution environments, differential privacy, zero knowledge proofs, bloom filter.

Question 13 - Can you elaborate a bit more on federated learning, trusted execution environments and zero knowledge proofs?

- Federated learning: the data does not come together, but the analysis or algorithm 'travels' to the data rather than the reverse. So, the model is updated by the party and then travels to the next one, to put it very simply.

- Trusted execution environments: is hardware based. No other applications can run, so it's a trusted environment to run some tasks.
- Zero knowledge proofs: allow you to prove that you know or something, but don't reveal anything more than that.

Question 14 -Looking towards the future, what does this mean for the strength of the math or cryptography protecting the data? Does the protection evolve along with the tools to break such technologies and with the increase in data?

- Of course, the GDPR also refers to the state of the art in instances, so you constantly have to update your techniques. Secret sharing is information theoretically safe still in the future, assuming that one communicates safely. But many schemes based on computational assumptions need to be updated with new technology advances such as quantum computing.
- What could be helpful is more advice from governmental agencies or DPA's on how we should view technologies. Of course, the GDPR is technology neutral and that is good, but still the guidelines from actors such as DPA's could be more precise. For example, on how technologies can be used and how they can help exactly, in the short term. PET's technologies are challenging from a legal and technical point of view, in bringing the two together.

7.1.10 De Wolf

Interview 20-04-2022
P.P. de Wolf – Statistics Netherlands (topic: SDC and statistical data)

Question 1 – In one of your papers you outline how the abundance of data poses new problems for statistical disclosure control. You outline various ways for an National Statistical Institution to deal with these disclosure risks (1- NSI may protect its open data taking into account other data sources, 2- NSI may act as an open-data provider for all public data 3- NSI may add legal conditions to the use of its open data that prohibit disclosure of individual data also when it is achieved by combining the NSI data with other public or private data sources). Can you elaborate a bit more on the different scenarios? Or more simply put, what are the main issues for a NSI in a world that is increasingly datafied?

- It is related to some of the ways in which we provide access to micro data. We make public use files, which are intended for the public at large, so for informational or educational purposes. Those micro data files have to be very protected in terms of statistical disclosure control (SDC). There should be no, or virtually no, possibilities of identifying persons in those data. If you can still identify people, then at least there should be no sensitive information in that dataset. So, we never have sensitive information in public use files. What is sensitive is not just determined in terms of the GDPR special data categories, but the Statistic Netherlands (SN) also takes into account what is societally seen as more sensitive information, such as information related to income. Of course, what is deemed sensitive is very context dependent, it can change over time and can differ per country, etc. So that is one scenario. Another scenario is where we have the tabular data on Statline, which is open data in some sense. We protect those data against disclosure even if combined with other information. Of course, it is not possible to take into account all data and all sources that are available, but the subject matter people at SN usually know what publications are out there, so at least we take those into account. A third scenario for giving access to micro data would be that you have some legal restrictions. For example, determining which actors have access to which data (e.g. researchers from specific institutes) and they can sign contracts with specific conditions and laying down consequences of disclosing information, or only on-site access for certain information where the outcome of the analysis is also checked for confidentiality. This is also part of the CBS Act.

Follow up question: in the paper in which you describe such scenarios you describe that there should be a mix of these different measures/scenarios? Is that correct?

- What we mean by that is there is an interaction between the level of SDC and the level of legal protection: legal protection e.g. by means of contractual agreements may lead to a different attacker scenario where the needed SDC methods are less stringent. It is a mix in the sense that we can change those levels for each publication or each access we provide. It is also a mixture of approaches in that you can see the various scenarios in the output that we provide. So we have public use files, secure use files, scientific use files: public use files are for the general public, secure and scientific use files are both for established researchers Secure use files are only for onsite or remote access, scientific use files can also be given access through DANS.

Question 2 – You are talking about giving access to micro data, and you mentioned scientific analysis. Does scientific analysis have a confined meaning to you/what are the restrictions? (e.g. specific people have access)

- Only employees from established research institutes are allowed to get access, those institutes and conditions for those are mentioned in the CBS Act. Once the researcher has access, both the researcher and the institute have to sign a confidentiality agreement and it also comes with specific rules (e.g. the research institute can be faced with the consequences of misbehavior from the researcher vis a vis the data). For each project we make have an intake to discuss what the purpose is, if we have the microdata etc. and only the data that is needed for that specific research

is provided (data limitation). If they have several projects, they also get separate environments so there is no data linking between environments. Whatever they produce in terms of results has to become publicly available, after it has been checked by CBS staff for confidentiality issues.

Question 3 – In our study we look at the fluidity of types of data. In your work you talk about how even when data are aggregated, they can still contain some personal information. Is there a technical measure in that sense to determine at which level data are anonymous?

- We try to not refer to anonymous data, as in our field data are never really anonymous data. But perhaps that is more a matter of definition, of what you would consider anonymous or not. We have a handbook with rules which specifies what each of our publications should comply with. E.g. if you have a frequency count table there should be a minimum number of people in a cell, or if you have quantitative magnitude tables there are rules on dominance as there should not be one entity that dominates the cell value. We also check for group disclosure, so if you have an identifiable group and everyone is scored on the same category everyone has a value on that category, thus if the group is too small it is not allowed to be published. It depends on the threshold, in the end statistics is about publishing data on groups of people, so the threshold determines what can be published or not. We do not give out the threshold values so as to protect the procedure.

Follow up question: does that also apply to other NSI's in other countries, that you do not make public the thresholds?

- Yes, also not in other EU countries. But each country is allowed to set their own thresholds. We, SN, also provide courses at Eurostat on SDC, so other countries also take over some of our ideas on SDC as we SN have been at the forefront of SDC in Europe in quite some years.

Follow up question: To determine which factors/data can provided, does that depend heavily on the context? (E.g. data about people in a small town might be more easily identifiable than data bout people in a bigger city)

- Yes, the context is important. We always look at the target population, e.g. the target population is not always everyone in the Netherlands, it usually concerns specific groups and they differ in size. We also take into account different scenarios. So, for example in public use files you should not be able to identify people in the dataset (high threshold), while for scientific use files because these are only available to specific people under specific conditions you want to protect against spontaneous recognition (the researcher has no intention to disclose something but could still recognize someone in the data). To do that we look at three dimensional combinations, so three values combined. This resembles k-anonymity a bit, so there should be at least a certain number of records in your dataset. For spontaneous recognition there should be several combinations of a key of three variables and all these combinations need to appear multiple times in the target population. As the NSI we have the advantage that because we also have a lot of administrative data we usually know the number of people in the target population for (some of) these combinations.

Question 4 – In one of your papers you describe that there are increasingly new challenges for NSI's in the big data era. What are some specific challenges that you are currently still facing?

- The major challenge is that there is still increasingly more open data available also from other organizations. How do you find all the open data that is available that you should take into account and then once you are aware if its existence how do you take that information into account. If you combine more and more information, it becomes increasingly difficult to protect against disclosure. Big data is a challenge as it is unstructured data, it is not always clear what the population is behind the data. So you don't always know what the entity is that the data is about and if it is unique in a population or not. And it is usually more event data rather than

observing a specific unit, so you have to translate what that data tells you about a unit, e.g. whether it is personal data or not.

Question 5 – We have from a legal perspective the notions of singling out, linkability and information inference. In differential privacy the notion of singling out has a specific meaning (guessing or assuming if someone is in a dataset to get information, rather than identifying someone in the data). Do you think this broader approach to singling out has merit or should we rather focus on the identifiability aspect?

- There are still discussions about whether we can use differential privacy in official statistics. It is being used in some private sector organizations, but of course for statistical purposes we have different aims for which we are using the data and a different balance with utility of data. For official statistics it is difficult to use because you quickly lose a lot of utility of the data, while we need to protect very accurate estimates of what is happening in society. We distinguish between identity disclosure and attribute disclosure. Often identity disclosure is the first step to attribution disclosure, so first you would identify a person and then you derive information about that person. But it is not always necessarily the case, you don't always get attribute disclosure. Differential privacy is more of a membership disclosure, it tells you whether you can or cannot say whether a person is in the dataset. It is the ratio of two probabilities or estimates, the dataset with and without the particular record, but you need both of these estimates. The notion of the privacy budget is very nice, especially from a mathematical point of view.

Question 6 – Staying on the topic of identity and attribute disclosure: something that you mention in one of your papers is that SDC methods aim currently at reducing the risk of identity disclosure, but maybe disclosure control should focus more on attribute disclosure as well. Can you tell us more about that?

- What you try to prevent is that you would be able to derive information about a person, whether you can identify that person is another question. It is more important that you target the attribute disclosure, such as disclosing sensitive information about a person. Whether you actually identify the person in the dataset is not always the case and is only part of the problem. If you prevent identification, it partially presents attribute disclosure, but not completely.

Question 7 – Do you think we should let go of the notion of linkage because we do not know all the information that is out there?

- Disclosure is usually linked to an attacker scenario. For example for scientific and secure use files you will not assume that the researcher will try to link their own dataset with the data from SN, while if it is a public use file that is a very realistic scenario. The linkage scenario is one of the scenario's that you should take into account. The measures that we take are chosen with a specific scenario in mind, you try to prevent that specific scenario. Of course, you cannot protect against all possible scenarios, there can be very out of the box unforeseen scenarios in some cases. The GDPR also asks what is reasonably likely only.

Follow question: what do you take into account for determining what is reasonably likely?

- What we take into account is other publications in the field that you are publishing in, you take those into account. Because it is likely or reasonable that people will use those. From the research perspective we also look at publications, e.g. ones that describe new attacker scenarios, and reflect on what those mean for us. Sometimes there is also an extra check if necessary, where we would let a SDC expert look at our publication to test whether they could derive certain information from that publication to test how secure it is. Differential privacy e.g. is said not to be dependent on a specific attacker scenario, but that can be nuanced because their scenario is mostly a membership attack.

Question 8 – To what extent do attacker scenarios change? So are they dependent on the state of the art of what is possible now and they change in the near future, or are they more consistent over time?

- You could say that the current rules that we use are based on relatively old attacker scenarios. But you can also distinguish between attacker scenarios to explain it better: one is what are you trying to disclose (e.g. identity disclosure or other factors), and the other side of the scenario is how are you going to disclose that information (e.g. do I have my own data or am I only guessing that someone is present in the data). The latter changes much more over time than the former, the methods that are available to disclose are the changing component in the scenario (e.g. better computing power). And at the research department we keep up with publications on attacks and assess if we need to take that information into account, in addition we also participate in meetings on the state of the art of SDC and attacker scenarios.

Question 9 – What is the role of quantum computing? Will this be a concern for NSIs?

- I am not an expert on quantum computing but from what I know the power of quantum computing is still quite low. Quantum computers will only be very beneficial for specific situations, for the practices of SN it will probably not change much for us in the near future. It is more relevant for actors working with a lot of encryption as protection, while of course we focus more on generalizations, adding noise, etc.

Question 10 – Can you explain to us a little bit more how different methods work together, so how do SDC and Privacy Preserving Data Sharing & Privacy Preserving Analysis (within Privacy Preserving Techniques) come together?

- PPT focuses more on the input side, you share information or apply an analysis without seeing the actual data. SDC targets the results of the analysis, what information can you gather from the results. In that sense they complement each other. You often need both, they are two methods. So technically I would not phrase SDC as a PPT.

Question 11 – In your paper you mention many different techniques of PPT. What are the promising techniques for statistical data or is that difficult to say in such a general way?

- That is difficult to answer, at the moment we have proof of concepts. PPT are more applied by the IT department of SN, they look at the question of whether we can apply that technique. In the methodology department we assess whether we actually need those techniques and what are the benefits of those techniques, and that depends more on the situation. It depends mostly on the number of parties in addition to SN working together. PPT work on the input side, but of course sometimes the output of one party is the input of another party, so that blurs the distinction a bit (because it depends from which angle you are looking at).

Question 12 – In your paper you describe ways in which we do not have to rely on trusted third parties. I wonder why that is, so why should we also look for alternatives that do not include trusted third parties?

- The assumption is that a trusted third party knows everything, thus you need a lot of trust from all other members in the multi-party situation. So when that trust is not fully there you need alternatives. To give an example: In a proof of concept that we did we had a multi-party computation, and everything was done with homomorphic encryption. But then for the results you still need SDC by the NSI. So, here the trusted third party does not do the analysis but only checks the output. So that is another way to go about it, it is just a smaller role for the trusted third party.

7.2 Workshop report

Date 3-3-2022

Participants:

1. Ashur (TuE)
2. Attoresi (EDPS)
3. Cesar Augusto Fontanillo Lopez (KU Leuven)
4. Barbera (BitnessWise)
5. Bholasing (VU)
6. Binns (Oxford University)
7. Bodea (TNO)
8. Demeyer (Waag)
9. Dercksen (Radboud University Nijmegen)
10. Van Eijk (Future of Privacy Forum),
11. Evers (Raad van State)
12. Koops (Tilburg University)
13. Klos (Autoriteit Persoonsgegevens)
14. Klous (KPMG)
15. Kusters (Viacryp)
16. Mommers (CBS)
17. Oosterbeek (Min OCW)
18. Preneel (KU Leuven)
19. Van Schendel (Tilburg University)
20. Sheikhalishahi (OU)
21. Van der Sloot (Tilburg University)
22. Stevens (CWI)
23. Walraven (Min OCW)
24. De Wolf (CBS)

Discussion points per topic:

Anonymous data and aggregation of data:

- Complexity of the terminology: what do the terms anonymization and aggregation mean? For example, if we speak about anonymity in terms of not being able to identify individuals, from a point of view of the technology this chance is almost never zero. The definition in itself is valid, as if there is only the possibility for one actor to identify individuals in the data, it should be treated as personal data. However, the technical reality of it being very difficult to achieve that true anonymity complicates matters.
- Black or white approach: The black or white approach of anonymity under the GDPR can be a demotivating factor for data controllers/processors, as it can be difficult to achieve true anonymity and there are no rewards (in terms of exemptions from legal obligations for example) for achieving a certain degree of anonymity. There is a misalliance between the technical/mathematical perspective, in which anonymization is not truly possible, and the legal regime. Previous to the GDPR the combination of data also led to de-identify problems, but the black and white approach to the scope of the GDPR in terms of anonymous or non-anonymous data does not help matters. The binary approach of the GDPR has clear limitations in its terminology towards the means of identification as 'reasonably likely'. Also in a temporal sense there is an evanescence in anonymization. There could be other systems possible, such as going more towards best effort or malice versus error-based systems. limitations about thinking about black and white regime, binary approach in the law.

- Contextual norms: would it be possible to see the norms of the GDPR as contextual norms? The context in which data is situated is an important factor for regulation. Different actors have different resources in terms of anonymizing and de-anonymizing data. These differences could be taken into account in addressees of norms. There are also dimensions of space and time that influence the anonymity of data and applicability of the GDPR.
- Legitimate interests: When it concerns aggregated data it is also important to remember the (societal)value of publishing and sharing it. Therefore, next to norms we could do with more pragmatic requirements or guidelines so that we can take legitimate interests into account.
- Level of the data being regulated: in terms of anonymization and aggregation there is a big difference in looking at the data from a record level or a more aggregate level. On a record level, it is almost impossible to speak of anonymous data where on the aggregate level there are many more opportunities to protect individuals from identification.
- Purpose of the law: we have to keep in mind the purpose of the law to assess which data to protect and how and balance the different interests that come into play, such as individual versus societal interests. In this regard there is a role for data ethics and data governance to play in addition to the GDPR.
- Privacy/identity preserving techniques: over the years, the popular techniques to presumably protect the identify of individuals, such as varieties of differential privacy and k-anonymity, have failed in providing absolute protection. The main reason for this is other types of information that are nowadays readily available, which the privacy preserving techniques do not fully take into account.
- Location data: a type of data that seems the most challenging to anonymize, and which is surrounded with misconceptions about that fact, is location data.
- Towards a greyer approach: if we were to move from a black and white regime to a more grey approach, there are numerous factors to take into account. We could argue that the GDPR applies to almost any dataset. So rather than focusing on the scope of the GDPR as such, an option could be to look at grey zones or a 'GDPR light regime' for some categories of data. One such factors that could be relevant is the assessment of harm and from whose perspective we should view the harm, which is important in assessing which data to protect from identifiability and to what extent. Another factor could be the privacy effort that has been made regarding the data. However, for all of the possible boundaries that we can draw, legal certainty is important.
- Precautionary principle: another possible solution would be a form of a precautionary principle, under which you invite controllers to behave as if data are presumably anonymized. It may be possible that the GDPR does not apply legally, but that it does provide tools for accountability.

Pseudonymous data:

- The definition/concept: Pseudonymous data originally meant 'data that is not directly identifiable'. The definition of pseudonymous data as it currently stands under the GDPR can be challenged, the category of pseudonymous data is meanly seen as a technical security measure.
- The choice of the legislator for pseudonymization: Why is pseudonymous data the technical measure that is chosen for exceptions in the GDPR? Obviously, there are also other technical means to diminish risk, pseudonymization is but one way to diminish risk. We need to consider whether pseudonymization is the best boundary marker for a lighter legal regime, or as it currently stands, if pseudonymization should indeed be the only means that receives such a status in the GDPR. On the other hand, it can be argued that it is incredibly difficult to come up with perfect boundary markers for a lighter regime or perfect intermediate categories of data between anonymous and fully identifiable personal data.
- Pseudonymization techniques: Hashing is often claimed to create pseudonymization, however in reality this protection is often easily broken. Another technique that could be more prominent

in the future is the use of synthetic data. For pseudonymization techniques it is important to assess whether they actually offer said protection to prevent abuse of legal exceptions. At least for every technique the base line is that the protective technique should be conducted properly, if applied properly it can offer protection, if not then it can lead to abuse of said technology.

- Purpose of pseudonymization: we can conclude that the main purpose of pseudonymization is to reduce risk. The question is how to assess the risk: For example, should we include other measures that can reduce risks for data subjects? Or do we assess the risk vis a vis a dataset or should we look at the system as a whole? And do we distinguish between different risks, for example distinguish between abuse or malice and other types of re-identification.

Sensitive data:

- Categories of sensitive data: The current categories are perhaps not necessarily the most sensitive ones. We could for example consider, financial information, location data, poverty, meta data and so forth. The list should not be exhausting. The use of categories can be helpful to flag the risk in processing operations. In addition, we also need to consider how to treat data that are frequently used to derive sensitive data.
- Feasibility of the approach: it is extremely difficult to choose the categories of data that should be deemed sensitive. Nonetheless there seems to be no clear alternative. A possible alternative could be a very abstract approach without categories but looking at the sensitive of the processing as a whole (if sensitiveness is a little, intermediate, or high) and categorize it accordingly. However, that still leaves open the question how to determine the different levels of sensitivity. We could for example borrow from mosaic privacy theories (where information combined gives a certain picture of an aspect of someone's life). But the law does not have the tools or guidelines to categorize that, it would require lower-level regulation or codes. The same goes for the current approach of the GDPR which speaks of a significant risk but leaves legal uncertainty as to what constitutes a significant risk.

Communications data – metadata

- The distinction between the two in law: meta data reveal a lot of information about a person. From a historical perspective we awarded different protection to the content of conversations in constitutions and different protection to other data pertaining the conversation, but this distinction no longer seems very relevant. Both type of data will be personal data under the GDPR anyway.

7.3 Provisions in GDPR

	Recital	Article
Personal data	(14) The protection afforded by this Regulation should apply to natural persons, whatever their nationality or place of residence, in relation to the processing of their personal data. This Regulation does not cover the processing of personal data which concerns legal persons and in particular undertakings established as legal persons, including the name and the form of the legal person and the contact details of the legal person.	4(1) 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;
	(27) This Regulation does not apply to the personal data of deceased persons. Member States may provide for rules regarding the processing of personal data of deceased persons.	
Anonymous data	(26) The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.	
Statistical/aggregated data	(50) The processing of personal data for purposes other than those for which the personal data were initially collected should be allowed only where the processing is compatible with the purposes for which the personal data were initially collected. In such a case, no legal basis separate from that which allowed the collection of the personal data is required. If the processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller, Union or Member State law may determine and specify the tasks and purposes for which the further processing should be regarded as compatible and lawful. Further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes should be considered to be compatible lawful processing operations. The legal basis provided by Union or Member State law for the processing of personal data may also provide a legal basis for further processing. In order to ascertain whether a purpose of further processing is compatible with the purpose for which the personal data are initially collected, the controller, after having met all the requirements for the lawfulness of the original processing, should take into account, inter alia: any link between those purposes and the purposes of the intended further processing; the context in which the personal data have been collected, in particular the reasonable expectations of data subjects based on their relationship with the controller as to their further use; the nature of the personal data; the consequences of the intended further processing for data subjects; and the existence of appropriate safeguards in both the original and intended further processing operations. Where the data subject has given consent or the processing is based on Union or Member State law which constitutes a necessary and proportionate measure in a democratic society to safeguard, in particular, important objectives of general public interest, the controller should be allowed to further process the personal data irrespective of the compatibility of the purposes. In any case, the application of the principles set out in this Regulation and in particular the information of the data subject on those other purposes and on his or her rights including the right to object, should be ensured. Indicating possible criminal acts or threats to public security by the controller and transmitting the relevant personal data in individual cases or in several cases relating to the same criminal act or threats to public security to a competent authority should be regarded as being in the legitimate interest pursued by the controller. However, such transmission in the legitimate interest of the controller or further processing of personal	5.1(b) collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation');

	data should be prohibited if the processing is not compatible with a legal, professional or other binding obligation of secrecy.	
		5.1(e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) subject to implementation of the appropriate technical and organisational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject ('storage limitation');
	(52) Derogating from the prohibition on processing special categories of personal data should also be allowed when provided for in Union or Member State law and subject to suitable safeguards, so as to protect personal data and other fundamental rights, where it is in the public interest to do so, in particular processing personal data in the field of employment law, social protection law including pensions and for health security, monitoring and alert purposes, the prevention or control of communicable diseases and other serious threats to health. Such a derogation may be made for health purposes, including public health and the management of health-care services, especially in order to ensure the quality and cost-effectiveness of the procedures used for settling claims for benefits and services in the health insurance system, or for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes. A derogation should also allow the processing of such personal data where necessary for the establishment, exercise or defence of legal claims, whether in court proceedings or in an administrative or out-of-court procedure.	9.2 Paragraph 1 shall not apply if one of the following applies: (j) processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject.
	(53) Special categories of personal data which merit higher protection should be processed for health-related purposes only where necessary to achieve those purposes for the benefit of natural persons and society as a whole, in particular in the context of the management of health or social care services and systems, including processing by the management and central national health authorities of such data for the purpose of quality control, management information and the general national and local supervision of the health or social care system, and ensuring continuity of health or social care and cross-border healthcare or health security, monitoring and alert purposes, or for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, based on Union or Member State law which has to meet an objective of public interest, as well as for studies conducted in the public interest in the area of public health. Therefore, this Regulation should provide for harmonised conditions for the processing of special categories of personal data concerning health, in respect of specific needs, in particular where the processing of such data is carried out for certain health-related purposes by persons subject to a legal obligation of professional secrecy. Union or Member State law should provide for specific and suitable measures so as to protect the fundamental rights and the personal data of natural persons. Member States should be allowed to maintain or introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or data concerning health. However, this should not hamper the free flow of personal data within the Union when those conditions apply to cross-border processing of such data.	
	(62) However, it is not necessary to impose the obligation to provide information where the data subject already possesses the information, where the recording or disclosure of the personal data is expressly laid down by law or where the provision of information to the data subject proves to be impossible or would involve a disproportionate effort. The latter could in particular be the case where processing is carried out for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes. In that regard, the number of data subjects, the age of the data and any appropriate safeguards adopted should be taken into consideration.	14.5 Paragraphs 1 to 4 shall not apply where and insofar as: (b) the provision of such information proves impossible or would involve a disproportionate effort, in particular for processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, subject to the conditions and safeguards referred to in Article 89(1) or in so far as the obligation referred to in paragraph 1 of this Article is likely to render impossible or seriously impair the achievement of the objectives of that processing. In such cases the controller shall take appropriate measures to protect the data subject's rights and freedoms and legitimate interests, including making the information publicly available;
	(65) A data subject should have the right to have personal data concerning him or her rectified and a 'right to be forgotten' where the retention of such data infringes this Regulation or Union or Member State law to which the controller is subject. In particular, a data subject should have the right to have his or her personal data erased and no longer processed where the	17.3 Paragraphs 1 and 2 shall not apply to the extent that processing is necessary: d) for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) in so far

	personal data are no longer necessary in relation to the purposes for which they are collected or otherwise processed, where a data subject has withdrawn his or her consent or objects to the processing of personal data concerning him or her, or where the processing of his or her personal data does not otherwise comply with this Regulation. That right is relevant in particular where the data subject has given his or her consent as a child and is not fully aware of the risks involved by the processing, and later wants to remove such personal data, especially on the internet. The data subject should be able to exercise that right notwithstanding the fact that he or she is no longer a child. However, the further retention of the personal data should be lawful where it is necessary, for exercising the right of freedom of expression and information, for compliance with a legal obligation, for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller, on the grounds of public interest in the area of public health, for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, or for the establishment, exercise or defence of legal claims.	as the right referred to in paragraph 1 is likely to render impossible or seriously impair the achievement of the objectives of that processing; or
	(71) The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention. Such processing includes 'profiling' that consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her. However, decision-making based on such processing, including profiling, should be allowed where expressly authorised by Union or Member State law to which the controller is subject, including for fraud and tax-evasion monitoring and prevention purposes conducted in accordance with the regulations, standards and recommendations of Union institutions or national oversight bodies and to ensure the security and reliability of a service provided by the controller, or necessary for the entering or performance of a contract between the data subject and a controller, or when the data subject has given his or her explicit consent. In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision. Such measure should not concern a child. In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect. Automated decision-making and profiling based on special categories of personal data should be allowed only under specific conditions.	21.6 Where personal data are processed for scientific or historical research purposes or statistical purposes pursuant to Article 89(1), the data subject, on grounds relating to his or her particular situation, shall have the right to object to processing of personal data concerning him or her, unless the processing is necessary for the performance of a task carried out for reasons of public interest.
	(113) Transfers which can be qualified as not repetitive and that only concern a limited number of data subjects, could also be possible for the purposes of the compelling legitimate interests pursued by the controller, when those interests are not overridden by the interests or rights and freedoms of the data subject and when the controller has assessed all the circumstances surrounding the data transfer. The controller should give particular consideration to the nature of the personal data, the purpose and duration of the proposed processing operation or operations, as well as the situation in the country of origin, the third country and the country of final destination, and should provide suitable safeguards to protect fundamental rights and freedoms of natural persons with regard to the processing of their personal data. Such transfers should be possible only in residual cases where none of the other grounds for transfer are applicable. For scientific or historical research purposes or statistical purposes, the legitimate expectations of society for an increase of knowledge should be taken into consideration. The controller should inform the supervisory authority and the data subject about the transfer.	
	(156) The processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes should be subject to appropriate safeguards for the rights and freedoms of the data subject pursuant to this Regulation. Those safeguards should	*89* 1.Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards, in accordance with this

	ensure that technical and organisational measures are in place in order to ensure, in particular, the principle of data minimisation. The further processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes is to be carried out when the controller has assessed the feasibility to fulfil those purposes by processing data which do not permit or no longer permit the identification of data subjects, provided that appropriate safeguards exist (such as, for instance, pseudonymisation of the data). Member States should provide for appropriate safeguards for the processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes. Member States should be authorised to provide, under specific conditions and subject to appropriate safeguards for data subjects, specifications and derogations with regard to the information requirements and rights to rectification, to erasure, to be forgotten, to restriction of processing, to data portability, and to object when processing personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes. The conditions and safeguards in question may entail specific procedures for data subjects to exercise those rights if this is appropriate in the light of the purposes sought by the specific processing along with technical and organisational measures aimed at minimising the processing of personal data in pursuance of the proportionality and necessity principles. The processing of personal data for scientific purposes should also comply with other relevant legislation such as on clinical trials.	Regulation, for the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner. 2.Where personal data are processed for scientific or historical research purposes or statistical purposes, Union or Member State law may provide for derogations from the rights referred to in Articles 15, 16, 18 and 21 subject to the conditions and safeguards referred to in paragraph 1 of this Article in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes. 3.Where personal data are processed for archiving purposes in the public interest, Union or Member State law may provide for derogations from the rights referred to in Articles 15, 16, 18, 19, 20 and 21 subject to the conditions and safeguards referred to in paragraph 1 of this Article in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes. 4.Where processing referred to in paragraphs 2 and 3 serves at the same time another purpose, the derogations shall apply only to processing for the purposes referred to in those paragraphs.
	(157) By coupling information from registries, researchers can obtain new knowledge of great value with regard to widespread medical conditions such as cardiovascular disease, cancer and depression. On the basis of registries, research results can be enhanced, as they draw on a larger population. Within social science, research on the basis of registries enables researchers to obtain essential knowledge about the long-term correlation of a number of social conditions such as unemployment and education with other life conditions. Research results obtained through registries provide solid, high-quality knowledge which can provide the basis for the formulation and implementation of knowledge-based policy, improve the quality of life for a number of people and improve the efficiency of social services. In order to facilitate scientific research, personal data can be processed for scientific research purposes, subject to appropriate conditions and safeguards set out in Union or Member State law.	
	(162) Where personal data are processed for statistical purposes, this Regulation should apply to that processing. Union or Member State law should, within the limits of this Regulation, determine statistical content, control of access, specifications for the processing of personal data for statistical purposes and appropriate measures to safeguard the rights and freedoms of the data subject and for ensuring statistical confidentiality. Statistical purposes mean any operation of collection and the processing of personal data necessary for statistical surveys or for the production of statistical results. Those statistical results may further be used for different purposes, including a scientific research purpose. The statistical purpose implies that the result of processing for statistical purposes is not personal data, but aggregate data, and that this result or the personal data are not used in support of measures or decisions regarding any particular natural person.	
	(163) The confidential information which the Union and national statistical authorities collect for the production of official European and official national statistics should be protected. European statistics should be developed, produced and disseminated in accordance with the statistical principles as set out in Article 338(2) TFEU, while national statistics should also comply with Member State law. Regulation (EC) No 223/2009 of the European Parliament and of the Council (2) provides further specifications on statistical confidentiality for European statistics.	
Profiling	(24) The processing of personal data of data subjects who are in the Union by a controller or processor not established in the Union should also be subject to this Regulation when it is related to the monitoring of the behaviour of such data subjects in so far as their behaviour takes place within the Union. In order to determine whether a processing activity can be considered to monitor the behaviour of data subjects, it should be	4 (4) 'profiling' means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health,

	ascertained whether natural persons are tracked on the internet including potential subsequent use of personal data processing techniques which consist of profiling a natural person, particularly in order to take decisions concerning her or him or for analysing or predicting her or his personal preferences, behaviours and attitudes.	personal preferences, interests, reliability, behaviour, location or movements;
	(38) Children merit specific protection with regard to their personal data, as they may be less aware of the risks, consequences and safeguards concerned and their rights in relation to the processing of personal data. Such specific protection should, in particular, apply to the use of personal data of children for the purposes of marketing or creating personality or user profiles and the collection of personal data with regard to children when using services offered directly to a child. The consent of the holder of parental responsibility should not be necessary in the context of preventive or counselling services offered directly to a child.	
	(60) The principles of fair and transparent processing require that the data subject be informed of the existence of the processing operation and its purposes. The controller should provide the data subject with any further information necessary to ensure fair and transparent processing taking into account the specific circumstances and context in which the personal data are processed. Furthermore, the data subject should be informed of the existence of profiling and the consequences of such profiling. Where the personal data are collected from the data subject, the data subject should also be informed whether he or she is obliged to provide the personal data and of the consequences, where he or she does not provide such data. That information may be provided in combination with standardised icons in order to give in an easily visible, intelligible and clearly legible manner, a meaningful overview of the intended processing. Where the icons are presented electronically, they should be machine-readable.	
	(63) A data subject should have the right of access to personal data which have been collected concerning him or her, and to exercise that right easily and at reasonable intervals, in order to be aware of, and verify, the lawfulness of the processing. This includes the right for data subjects to have access to data concerning their health, for example the data in their medical records containing information such as diagnoses, examination results, assessments by treating physicians and any treatment or interventions provided. Every data subject should therefore have the right to know and obtain communication in particular with regard to the purposes for which the personal data are processed, where possible the period for which the personal data are processed, the recipients of the personal data, the logic involved in any automatic personal data processing and, at least when based on profiling, the consequences of such processing. Where possible, the controller should be able to provide remote access to a secure system which would provide the data subject with direct access to his or her personal data. That right should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software. However, the result of those considerations should not be a refusal to provide all information to the data subject. Where the controller processes a large quantity of information concerning the data subject, the controller should be able to request that, before the information is delivered, the data subject specify the information or processing activities to which the request relates.	13.2 In addition to the information referred to in paragraph 1, the controller shall, at the time when personal data are obtained, provide the data subject with the following further information necessary to ensure fair and transparent processing: (f) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject. 14.2 In addition to the information referred to in paragraph 1, the controller shall provide the data subject with the following information necessary to ensure fair and transparent processing in respect of the data subject: g) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject. 15.1The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information: (h) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.
	(70) Where personal data are processed for the purposes of direct marketing, the data subject should have the right to object to such processing, including profiling to the extent that it is related to such direct marketing, whether with regard to initial or further processing, at any time and free of charge. That right should be explicitly brought to the attention of the data subject and presented clearly and separately from any other information.	21.1 The data subject shall have the right to object, on grounds relating to his or her particular situation, at any time to processing of personal data concerning him or her which is based on point (e) or (f) of Article 6(1), including profiling based on those provisions. The controller shall no longer process the personal data unless the controller demonstrates compelling legitimate grounds for the processing which override the interests, rights and freedoms of the data subject or for the establishment, exercise or defence of legal claims. 2.Where personal data are processed for direct marketing purposes, the data subject shall have the right to object at any time to processing of

		personal data concerning him or her for such marketing, which includes profiling to the extent that it is related to such direct marketing. 3.Where the data subject objects to processing for direct marketing purposes, the personal data shall no longer be processed for such purposes. 4.At the latest at the time of the first communication with the data subject, the right referred to in paragraphs 1 and 2 shall be explicitly brought to the attention of the data subject and shall be presented clearly and separately from any other information. 5.In the context of the use of information society services, and notwithstanding Directive 2002/58/EC, the data subject may exercise his or her right to object by automated means using technical specifications. 6.Where personal data are processed for scientific or historical research purposes or statistical purposes pursuant to Article 89(1), the data subject, on grounds relating to his or her particular situation, shall have the right to object to processing of personal data concerning him or her, unless the processing is necessary for the performance of a task carried out for reasons of public interest.
		22.1 The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her. 2.Paragraph 1 shall not apply if the decision: (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller; (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or (c) is based on the data subject's explicit consent. 3.In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision. 4.Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.
	(72) Profiling is subject to the rules of this Regulation governing the processing of personal data, such as the legal grounds for processing or data protection principles. The European Data Protection Board established by this Regulation (the 'Board') should be able to issue guidance in that context.	
	(73) Restrictions concerning specific principles and the rights of information, access to and rectification or erasure of personal data, the right to data portability, the right to object, decisions based on profiling, as well as the communication of a personal data breach to a data subject and certain related obligations of the controllers may be imposed by Union or Member State law, as far as necessary and proportionate in a democratic society to safeguard public security, including the protection of human life especially in response to natural or manmade disasters, the prevention, investigation and prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security, or of breaches of ethics for regulated professions, other important objectives of general public interest of the Union or of a Member State, in particular an important economic or financial interest of the Union or of a Member State, the keeping of public registers kept for reasons of general public interest, further processing of archived personal data to provide specific information related to the political behaviour under former totalitarian state regimes or the protection of the data subject or the rights and freedoms of others, including social protection, public health and humanitarian purposes. Those restrictions should be in accordance with the requirements set out in the Charter and in the European Convention for the Protection of Human Rights and Fundamental Freedoms.	

	(75) The risk to the rights and freedoms of natural persons, of varying likelihood and severity, may result from personal data processing which could lead to physical, material or non-material damage, in particular: where the processing may give rise to discrimination, identity theft or fraud, financial loss, damage to the reputation, loss of confidentiality of personal data protected by professional secrecy, unauthorised reversal of pseudonymisation, or any other significant economic or social disadvantage; where data subjects might be deprived of their rights and freedoms or prevented from exercising control over their personal data; where personal data are processed which reveal racial or ethnic origin, political opinions, religion or philosophical beliefs, trade union membership, and the processing of genetic data, data concerning health or data concerning sex life or criminal convictions and offences or related security measures; where personal aspects are evaluated, in particular analysing or predicting aspects concerning performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, in order to create or use personal profiles; where personal data of vulnerable natural persons, in particular of children, are processed; or where processing involves a large amount of personal data and affects a large number of data subjects.	
	(91) This should in particular apply to large-scale processing operations which aim to process a considerable amount of personal data at regional, national or supranational level and which could affect a large number of data subjects and which are likely to result in a high risk, for example, on account of their sensitivity, where in accordance with the achieved state of technological knowledge a new technology is used on a large scale as well as to other processing operations which result in a high risk to the rights and freedoms of data subjects, in particular where those operations render it more difficult for data subjects to exercise their rights. A data protection impact assessment should also be made where personal data are processed for taking decisions regarding specific natural persons following any systematic and extensive evaluation of personal aspects relating to natural persons based on profiling those data or following the processing of special categories of personal data, biometric data, or data on criminal convictions and offences or related security measures. A data protection impact assessment is equally required for monitoring publicly accessible areas on a large scale, especially when using optic-electronic devices or for any other operations where the competent supervisory authority considers that the processing is likely to result in a high risk to the rights and freedoms of data subjects, in particular because they prevent data subjects from exercising a right or using a service or a contract, or because they are carried out systematically on a large scale. The processing of personal data should not be considered to be on a large scale if the processing concerns personal data from patients or clients by an individual physician, other health care professional or lawyer. In such cases, a data protection impact assessment should not be mandatory.	35.3 A data protection impact assessment referred to in paragraph 1 shall in particular be required in the case of: (a) a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person;
		35.2 The binding corporate rules referred to in paragraph 1 shall specify at least: (e) the rights of data subjects in regard to processing and the means to exercise those rights, including the right not to be subject to decisions based solely on automated processing, including profiling in accordance with Article 22, the right to lodge a complaint with the competent supervisory authority and before the competent courts of the Member States in accordance with Article 79, and to obtain redress and, where appropriate, compensation for a breach of the binding corporate rules;
		70. 1.The Board shall ensure the consistent application of this Regulation. To that end, the Board shall, on its own initiative or, where relevant, at the request of the Commission, in particular: (f) issue guidelines, recommendations and best practices in accordance with point (e) of this paragraph for further specifying the criteria and conditions for decisions based on profiling pursuant to Article 22(2);
Open data/re -use	(153) Member States law should reconcile the rules governing freedom of expression and information, including journalistic, academic, artistic and or literary expression with the right to the protection of personal data pursuant to this Regulation. The processing of personal data solely for journalistic purposes, or for the purposes of academic, artistic or literary expression should be subject to derogations or exemptions from certain provisions of this Regulation if necessary to reconcile the right to the protection of personal data with the right to freedom of expression and information, as enshrined in Article 11 of the Charter. This should apply in particular to the processing of personal data in the audiovisual field and in	85.1 Member States shall by law reconcile the right to the protection of personal data pursuant to this Regulation with the right to freedom of expression and information, including processing for journalistic purposes and the purposes of academic, artistic or literary expression. 2.For processing carried out for journalistic purposes or the purpose of academic artistic or literary expression, Member States shall provide for exemptions or derogations from Chapter II

	news archives and press libraries. Therefore, Member States should adopt legislative measures which lay down the exemptions and derogations necessary for the purpose of balancing those fundamental rights. Member States should adopt such exemptions and derogations on general principles, the rights of the data subject, the controller and the processor, the transfer of personal data to third countries or international organisations, the independent supervisory authorities, cooperation and consistency, and specific data-processing situations. Where such exemptions or derogations differ from one Member State to another, the law of the Member State to which the controller is subject should apply. In order to take account of the importance of the right to freedom of expression in every democratic society, it is necessary to interpret notions relating to that freedom, such as journalism, broadly.	(principles), Chapter III (rights of the data subject), Chapter IV (controller and processor), Chapter V (transfer of personal data to third countries or international organisations), Chapter VI (independent supervisory authorities), Chapter VII (cooperation and consistency) and Chapter IX (specific data processing situations) if they are necessary to reconcile the right to the protection of personal data with the freedom of expression and information. 3.Each Member State shall notify to the Commission the provisions of its law which it has adopted pursuant to paragraph 2 and, without delay, any subsequent amendment law or amendment affecting them.
	(154) This Regulation allows the principle of public access to official documents to be taken into account when applying this Regulation. Public access to official documents may be considered to be in the public interest. Personal data in documents held by a public authority or a public body should be able to be publicly disclosed by that authority or body if the disclosure is provided for by Union or Member State law to which the public authority or public body is subject. Such laws should reconcile public access to official documents and the reuse of public sector information with the right to the protection of personal data and may therefore provide for the necessary reconciliation with the right to the protection of personal data pursuant to this Regulation. The reference to public authorities and bodies should in that context include all authorities or other bodies covered by Member State law on public access to documents. Directive 2003/98/EC of the European Parliament and of the Council leaves intact and in no way affects the level of protection of natural persons with regard to the processing of personal data under the provisions of Union and Member State law, and in particular does not alter the obligations and rights set out in this Regulation. In particular, that Directive should not apply to documents to which access is excluded or restricted by virtue of the access regimes on the grounds of protection of personal data, and parts of documents accessible by virtue of those regimes which contain personal data the re-use of which has been provided for by law as being incompatible with the law concerning the protection of natural persons with regard to the processing of personal data.	86 Personal data in official documents held by a public authority or a public body or a private body for the performance of a task carried out in the public interest may be disclosed by the authority or body in accordance with Union or Member State law to which the public authority or body is subject in order to reconcile public access to official documents with the right to the protection of personal data pursuant to this Regulation.
Pseudonymous data	(28) The application of pseudonymisation to personal data can reduce the risks to the data subjects concerned and help controllers and processors to meet their data-protection obligations. The explicit introduction of 'pseudonymisation' in this Regulation is not intended to preclude any other measures of data protection.	4(5) 'pseudonymisation' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;
	(29) In order to create incentives to apply pseudonymisation when processing personal data, measures of pseudonymisation should, whilst allowing general analysis, be possible within the same controller when that controller has taken technical and organisational measures necessary to ensure, for the processing concerned, that this Regulation is implemented, and that additional information for attributing the personal data to a specific data subject is kept separately. The controller processing the personal data should indicate the authorised persons within the same controller.	
	(30) Natural persons may be associated with online identifiers provided by their devices, applications, tools and protocols, such as internet protocol addresses, cookie identifiers or other identifiers such as radio frequency identification tags. This may leave traces which, in particular when combined with unique identifiers and other information received by the servers, may be used to create profiles of the natural persons and identify them.	
		6.4 Where the processing for a purpose other than that for which the personal data have been collected is not based on the data subject's consent or on a Union or Member State law which constitutes a necessary and proportionate measure in a democratic society to safeguard the objectives referred to in Article 23(1), the controller shall, in order to ascertain whether processing for another purpose is compatible with the purpose for which the personal data are initially collected, take into account, inter alia: (e) the existence of appropriate safeguards, which may include encryption or pseudonymisation.

	(75) The risk to the rights and freedoms of natural persons, of varying likelihood and severity, may result from personal data processing which could lead to physical, material or non-material damage, in particular: where the processing may give rise to discrimination, identity theft or fraud, financial loss, damage to the reputation, loss of confidentiality of personal data protected by professional secrecy, unauthorised reversal of pseudonymisation, or any other significant economic or social disadvantage; where data subjects might be deprived of their rights and freedoms or prevented from exercising control over their personal data; where personal data are processed which reveal racial or ethnic origin, political opinions, religion or philosophical beliefs, trade union membership, and the processing of genetic data, data concerning health or data concerning sex life or criminal convictions and offences or related security measures; where personal aspects are evaluated, in particular analysing or predicting aspects concerning performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, in order to create or use personal profiles; where personal data of vulnerable natural persons, in particular of children, are processed; or where processing involves a large amount of personal data and affects a large number of data subjects.	25.1Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.
	(78) The protection of the rights and freedoms of natural persons with regard to the processing of personal data require that appropriate technical and organisational measures be taken to ensure that the requirements of this Regulation are met. In order to be able to demonstrate compliance with this Regulation, the controller should adopt internal policies and implement measures which meet in particular the principles of data protection by design and data protection by default. Such measures could consist, inter alia, of minimising the processing of personal data, pseudonymising personal data as soon as possible, transparency with regard to the functions and processing of personal data, enabling the data subject to monitor the data processing, enabling the controller to create and improve security features. When developing, designing, selecting and using applications, services and products that are based on the processing of personal data or process personal data to fulfil their task, producers of the products, services and applications should be encouraged to take into account the right to data protection when developing and designing such products, services and applications and, with due regard to the state of the art, to make sure that controllers and processors are able to fulfil their data protection obligations. The principles of data protection by design and by default should also be taken into consideration in the context of public tenders.	32.1Taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons, the controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk, including inter alia as appropriate: (a) the pseudonymisation and encryption of personal data;
		40.2 Associations and other bodies representing categories of controllers or processors may prepare codes of conduct, or amend or extend such codes, for the purpose of specifying the application of this Regulation, such as with regard to: (d) the pseudonymisation of personal data;
	(85) A personal data breach may, if not addressed in an appropriate and timely manner, result in physical, material or non-material damage to natural persons such as loss of control over their personal data or limitation of their rights, discrimination, identity theft or fraud, financial loss, unauthorised reversal of pseudonymisation, damage to reputation, loss of confidentiality of personal data protected by professional secrecy or any other significant economic or social disadvantage to the natural person concerned. Therefore, as soon as the controller becomes aware that a personal data breach has occurred, the controller should notify the personal data breach to the supervisory authority without undue delay and, where feasible, not later than 72 hours after having become aware of it, unless the controller is able to demonstrate, in accordance with the accountability principle, that the personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons. Where such notification cannot be achieved within 72 hours, the reasons for the delay should accompany the notification and information may be provided in phases without undue further delay.	
	(156) The processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes should be subject to appropriate safeguards for the rights and freedoms of the data subject pursuant to this Regulation. Those safeguards should ensure that technical and organisational measures are in place in order to ensure, in particular, the principle of data minimisation. The further processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes is to be carried out when the controller has assessed the feasibility to fulfil those purposes by processing data which do not permit or no longer permit the identification of data subjects, provided that appropriate safeguards exist (such as, for instance, pseudonymisation of the data). Member States should provide for appropriate safeguards for the processing of personal data for archiving purposes in the public interest, scientific or historical	

	research purposes or statistical purposes. Member States should be authorised to provide, under specific conditions and subject to appropriate safeguards for data subjects, specifications and derogations with regard to the information requirements and rights to rectification, to erasure, to be forgotten, to restriction of processing, to data portability, and to object when processing personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes. The conditions and safeguards in question may entail specific procedures for data subjects to exercise those rights if this is appropriate in the light of the purposes sought by the specific processing along with technical and organisational measures aimed at minimising the processing of personal data in pursuance of the proportionality and necessity principles. The processing of personal data for scientific purposes should also comply with other relevant legislation such as on clinical trials.	
Sensitive personal data	(34) Genetic data should be defined as personal data relating to the inherited or acquired genetic characteristics of a natural person which result from the analysis of a biological sample from the natural person in question, in particular chromosomal, deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) analysis, or from the analysis of another element enabling equivalent information to be obtained.	4 (13) 'genetic data' means personal data relating to the inherited or acquired genetic characteristics of a natural person which give unique information about the physiology or the health of that natural person and which result, in particular, from an analysis of a biological sample from the natural person in question;
		4 (14)'biometric data' means personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data;
	(35) Personal data concerning health should include all data pertaining to the health status of a data subject which reveal information relating to the past, current or future physical or mental health status of the data subject. This includes information about the natural person collected in the course of the registration for, or the provision of, health care services as referred to in Directive 2011/24/EU of the European Parliament and of the Council (1) to that natural person; a number, symbol or particular assigned to a natural person to uniquely identify the natural person for health purposes; information derived from the testing or examination of a body part or bodily substance, including from genetic data and biological samples; and any information on, for example, a disease, disability, disease risk, medical history, clinical treatment or the physiological or biomedical state of the data subject independent of its source, for example from a physician or other health professional, a hospital, a medical device or an in vitro diagnostic test.	4 (15) 'data concerning health' means personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status;
	(51) Personal data which are, by their nature, particularly sensitive in relation to fundamental rights and freedoms merit specific protection as the context of their processing could create significant risks to the fundamental rights and freedoms. Those personal data should include personal data revealing racial or ethnic origin, whereby the use of the term 'racial origin' in this Regulation does not imply an acceptance by the Union of theories which attempt to determine the existence of separate human races. The processing of photographs should not systematically be considered to be processing of special categories of personal data as they are covered by the definition of biometric data only when processed through a specific technical means allowing the unique identification or authentication of a natural person. Such personal data should not be processed, unless processing is allowed in specific cases set out in this Regulation, taking into account that Member States law may lay down specific provisions on data protection in order to adapt the application of the rules of this Regulation for compliance with a legal obligation or for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller. In addition to the specific requirements for such processing, the general principles and other rules of this Regulation should apply, in particular as regards the conditions for lawful processing. Derogations from the general prohibition for processing such special categories of personal data should be explicitly provided, inter alia, where the data subject gives his or her explicit consent or in respect of specific needs in particular where the processing is carried out in the course of legitimate activities by certain associations or foundations the purpose of which is to permit the exercise of fundamental freedoms. (52) Derogating from the prohibition on processing special categories of personal data should also be allowed when provided for in Union or Member State law and subject to suitable safeguards, so as to protect personal data and other fundamental rights, where it is in the public interest to do so, in particular processing personal data in the field of employment law, social protection law including pensions and for health security, monitoring and alert purposes, the prevention or control of communicable diseases and other serious threats to health. Such a	9.1.Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited. 2.Paragraph 1 shall not apply if one of the following applies: (a) the data subject has given explicit consent to the processing of those personal data for one or more specified purposes, except where Union or Member State law provide that the prohibition referred to in paragraph 1 may not be lifted by the data subject; (b) processing is necessary for the purposes of carrying out the obligations and exercising specific rights of the controller or of the data subject in the field of employment and social security and social protection law in so far as it is authorised by Union or Member State law or a collective agreement pursuant to Member State law providing for appropriate safeguards for the fundamental rights and the interests of the data subject; (c) processing is necessary to protect the vital interests of the data subject or of another natural person where the data subject is physically or legally incapable of giving consent; (d) processing is carried out in the course of its legitimate activities with appropriate safeguards by a foundation, association or any other not-for-profit body with a political, philosophical, religious or trade union aim and on condition that the processing relates solely to the members or to former members of the body or to persons who have regular contact with it in connection with its

	derogation may be made for health purposes, including public health and the management of health-care services, especially in order to ensure the quality and cost-effectiveness of the procedures used for settling claims for benefits and services in the health insurance system, or for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes. A derogation should also allow the processing of such personal data where necessary for the establishment, exercise or defence of legal claims, whether in court proceedings or in an administrative or out-of-court procedure. (53) Special categories of personal data which merit higher protection should be processed for health-related purposes only where necessary to achieve those purposes for the benefit of natural persons and society as a whole, in particular in the context of the management of health or social care services and systems, including processing by the management and central national health authorities of such data for the purpose of quality control, management information and the general national and local supervision of the health or social care system, and ensuring continuity of health or social care and cross-border healthcare or health security, monitoring and alert purposes, or for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, based on Union or Member State law which has to meet an objective of public interest, as well as for studies conducted in the public interest in the area of public health. Therefore, this Regulation should provide for harmonised conditions for the processing of special categories of personal data concerning health, in respect of specific needs, in particular where the processing of such data is carried out for certain health-related purposes by persons subject to a legal obligation of professional secrecy. Union or Member State law should provide for specific and suitable measures so as to protect the fundamental rights and the personal data of natural persons. Member States should be allowed to maintain or introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or data concerning health. However, this should not hamper the free flow of personal data within the Union when those conditions apply to cross-border processing of such data. (54) The processing of special categories of personal data may be necessary for reasons of public interest in the areas of public health without consent of the data subject. Such processing should be subject to suitable and specific measures so as to protect the rights and freedoms of natural persons. In that context, 'public health' should be interpreted as defined in Regulation (EC) No 1338/2008 of the European Parliament and of the Council (1), namely all elements related to health, namely health status, including morbidity and disability, the determinants having an effect on that health status, health care needs, resources allocated to health care, the provision of, and universal access to, health care as well as health care expenditure and financing, and the causes of mortality. Such processing of data concerning health for reasons of public interest should not result in personal data being processed for other purposes by third parties such as employers or insurance and banking companies. (55) Moreover, the processing of personal data by official authorities for the purpose of achieving the aims, laid down by constitutional law or by international public law, of officially recognised religious associations, is carried out on grounds of public interest. (56) Where in the course of electoral activities, the operation of the democratic system in a Member State requires that political parties compile personal data on people's political opinions, the processing of such data may be permitted for reasons of public interest, provided that appropriate safeguards are established.	purposes and that the personal data are not disclosed outside that body without the consent of the data subjects; (e) processing relates to personal data which are manifestly made public by the data subject; (f) processing is necessary for the establishment, exercise or defence of legal claims or whenever courts are acting in their judicial capacity; (g) processing is necessary for reasons of substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject; (h) processing is necessary for the purposes of preventive or occupational medicine, for the assessment of the working capacity of the employee, medical diagnosis, the provision of health or social care or treatment or the management of health or social care systems and services on the basis of Union or Member State law or pursuant to contract with a health professional and subject to the conditions and safeguards referred to in paragraph 3; (i) processing is necessary for reasons of public interest in the area of public health, such as protecting against serious cross-border threats to health or ensuring high standards of quality and safety of health care and of medicinal products or medical devices, on the basis of Union or Member State law which provides for suitable and specific measures to safeguard the rights and freedoms of the data subject, in particular professional secrecy; j) processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject. 3.Personal data referred to in paragraph 1 may be processed for the purposes referred to in point (h) of paragraph 2 when those data are processed by or under the responsibility of a professional subject to the obligation of professional secrecy under Union or Member State law or rules established by national competent bodies or by another person also subject to an obligation of secrecy under Union or Member State law or rules established by national competent bodies. 4.Member States may maintain or introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or data concerning health.
		10 Processing of personal data relating to criminal convictions and offences or related security measures based on Article 6(1) shall be carried out only under the control of official authority or when the processing is authorised by Union or Member State law providing for appropriate safeguards for the rights and freedoms of data subjects. Any comprehensive register of criminal convictions shall be kept only under the control of official authority.