6
Regulating non-personal data in the age of Big Data

Bart van der Sloot

1. Introduction

The right to privacy has been included in national constitutions for centuries, with the distinction between the private and the public domain, the sanctity of the body and the secrecy of communication as core pillars of constitutional democracies around the world. After World War II, a number of human rights instruments were drawn up, such as the Universal Declaration of Human Rights,[1] the International Covenant on Civil and Political Human Rights[2] and the European Convention on Human Rights (ECHR).[3] Each of those included a specific article on the protection of the right to privacy, guaranteeing the respect for every individual's private life, family life, home and communication.

Informational privacy is an uncontroversial element of the right to privacy. Still, the right to privacy traditionally only covers information that is either private (falling under the protection of communicational secrecy) or sensitive (falling under the protection of private life). Privacy does not apply, or applies only to a limited extent, to the processing of public or insensitive data. In addition, when processing of personal information only has a minor effect on a person's private life, such would not be said to pass the threshold of the so-called *de minimis rule*, which is formalised in article 35 § 3 ECHR, providing that the European Court of Human Rights (ECtHR) should declare inadmissible any individual application if the applicant has not suffered a significant disadvantage.

Such logic has been applied in the context of data processing throughout the ECtHR's history. For example, when the European Commission of Human Rights (ECmHR) was faced with a person who felt that a photo taken of his vehicle violated his right to privacy, as protected under Article 8 ECHR, the Commission declared the claim inadmissible:

> Afin de déterminer dans des cas similaires l'étendue de la garantie accordée par l'article 8 contre les ingérences des autorités publiques, la Commission examine si la prise de photographies constitue une

> intrusion dans la sphère privée d'un individu (par exemple lorsqu'elles sont prises à son domicile), si les photographies se réfèrent à des événements d'ordre privé ou public, et si elles sont destinées à servir à un usage limité ou susceptibles d'être portées à la connaissance du public. En l'espèce, la Commission relève que la photographie dont se plaint le requérant a été prise sur la voie publique, alors qu'il circulait en voiture, dans un but de preuve et d'identification. Rien n'indique que la photographie ait été portée à la connaissance du public ni utilisée à d'autres fins que celle des poursuites dont le requérant a fait l'objet. Faisant application des critères exposés ci-dessus, la Commission arrive à la conclusion qu'il n'y a pas eu ingérence dans la vie privée du requérant.[4]

With the rise of databases in which large numbers of insensitive and public data were included, both the United States,[5] a number of European countries[6] and the Council of Europe[7] adopted specific data protection laws in the 1970s. The novelty of these instruments was that they provided protection to personal information that was not necessarily private or sensitive; they also covered the processing of data concerning, for example, the number of dogs owned by a person, her place of birth, whether she has a drivers licence, etc. The material scope of the right to data protection is not dependent on the existence of individual harm, but determined by the question of whether the data can be used to identify a person. 'Do you see that man there, with the black hat on, next to the streetlight', is considered personal data, even if identification does not have any effect on that person's private life or affects him in any significant way.

In contrast to human rights instruments, data protection regimes, such as the General Data Protection Regulation (GDPR), do not only provide protection to individual interests but typically aim at reconciling two interests, as is exemplified by Article 1 of that Regulation, specifying:

> 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 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.

The goal of the European Union (EU) data protection framework lies in its ambition to take away restrictions for data processing operations, while at the same time assuring a high level of protection for data subjects. One of the problems that existed before the EU data protection framework had been put in place – in the form of the predecessor of the GDPR, the 1995 Data

Protection Directive – was that each EU country adopted its own data protection standards. This hampered international data transfers and data-driven activities because an international organisation had to comply with multiple, sometimes conflicting, data protection requirements. By laying down a common data protection regime applicable throughout the EU, this problem is tackled, while at the same time providing protection to the interests of data subjects.

Two important developments have occurred since the 1970s.

First, the right to privacy and the right to data protection have grown apart. While initially, the right to data protection and the various data protection regimes were directly linked to and seen as a part of the right to privacy, at least within the EU, the right to data protection is now seen as a related though separate right. Inter alia, the Charter of Fundamental Rights of the European Union includes one article on the right to privacy and another on the right to data protection,[8] and while the Data Protection Directive from 1995 still held that the Directive provided protection in particular to the right to privacy,[9] the document that replaced it – the General Data Protection Regulation from 2016 – stresses that the instrument intends to provide further rules on the right to data protection, not the right to privacy, and has rephrased generally accepted terms such as 'privacy by design', 'privacy policy', 'privacy impact assessment' and 'privacy officer' to 'data protection by design', 'data protection policy', 'data protection impact assessment' and 'data protection officer'.[10]

Second, the material scope of data protection regimes has grown considerably over time. The Council of Europe adopted two Resolutions for data processing in 1973 and 1974, one for the private and one for the public sector, which defined 'personal information' simply as information relating to individuals (physical persons). Here, the individual and subjective element in the definition of personal data is still prominent. Already by 1981, however, in the subsequent Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, adopted by the Council of Europe, 'personal data' were defined as any information relating to an identified or identifiable individual.[11] An 'identifiable person' is an individual who at the present cannot be identified through the data, but in the future might be. This means that data that are not yet linked to an individual, but could be with relative ease in the future, already fall within the scope of the right to data protection. Under the subsequent Data Protection Directive, the scope of personal data was broadened, among others by including not only direct but also indirect identifiable data, and under the General Data Protection Regulation the definition was widened even further.

The reason for widening the scope of data protection regimes is that even data remotely connected to an individual can increasingly be used to gain insights on that person. While companies and governmental organizations used to be able to link data to a specific person only when they had direct identifiers, such as a name or an address, they are increasingly capable of connecting two or three indirect identifiers that in themselves do not refer to a specific person but might when combined. To accommodate this change, the concept of personal data has been broadened over time, to ensure that citizens are protected when organisations process these indirectly identifiable data. Some have suggested that the definition of personal data is now so broad that potentially all data could fall under its scope. Others have suggested that given that the technological capacities will only grow, it might be easier to simply let go of the notion of 'personal data' and regulate 'data' instead.[12]

Interestingly, the European Union has recently taken a contrary approach. Instead of broadening the scope of 'personal data' or accepting that the difference between personal data and non-personal data may be increasingly redundant, it has emphasised the polarity between the two types of data. In 2018, it adopted a Regulation on the transfer of non-personal data that only aims at stimulating cross-border data processing, without providing any form of protection to citizens. Where human rights documents aim at the protection of the interests of citizens and data protection regimes aim at reconciling the interests of citizens and the interests of organisations processing personal data, the Regulation on non-personal data only aims at protecting the interests of the latter. Article 1 of that Regulation specifies:

> This Regulation aims to ensure the free flow of data other than personal data within the Union by laying down rules relating to data localisation requirements, the availability of data to competent authorities and the porting of data for professional users.[13]

The material provisions of the Regulation do not aim at restricting or laying down conditions for the processing or transfer of non-personal data but in contrast, prohibit any type of restriction or limitation in national laws on the availability, transfer and processing of non-personal data.

This chapter will question the logic behind this choice. It will do so in three steps. First, it will discuss the tension between the current legal paradigm, which is grounded in a static conceptualisation of data, and the technological reality, in which the status of data is constantly in flux (Section 2). Second, it will argue why the boundaries between personal data and non-personal data, between metadata and content data, between anonymous and identifying data and between non-sensitive and sensitive data is increasingly difficult to draw (Section 3). Third, it will suggest why the logic underpinning the distinction between the various legal categories of data may no longer be valid in the age of Big Data (Section 4). Finally, the conclusion to this chapter will discuss the implications of these three arguments for the regulation of data (Section 5).

Two caveats are important. First, this chapter will mainly refer to and draw from legal definitions and examples from European legislative instruments. Although there may be some elements particular to the European situation, by and large, legal instruments around the world are based on much of the same premises and categorisations. Second, when referring to Big Data, this chapter will describe those processes in most simplistic terms, for example explaining how complex data analytics and computational modelling through self-learning algorithms work by referring to data analysis and profiling based on two or three data points; in reality, the number is usually closer to 200 or 300 data points that are interrelated in various ways. Although data analytics is consequently infinitely more complex than explained here, in essence it operates similarly to the examples provided here in basic terms for the sake of clarity and accessibility.

2. The status of data is unstable

Law works with definitions, categories and delineations. The moment something is defined, there is discussion about borderline cases. Does owning a tank fall under the right of the people to keep and bear arms as provided by the Second Amendment to the United States Constitution? Is shouting 'Fire, fire!' in a movie theatre covered by the freedom of expression? Should the Church of the Flying Spaghetti Monster be considered a religious institution for the purposes of the freedom of religion? These discussions are intensified when legal definitions and categories play a role in contexts in which rapidly evolving technologies and unforeseen applications emerge. When confronted with such questions and legal complexities, courts have generally adopted a flexible approach and opted for a broad understanding of subjective rights.

For example, the European Court of Human Rights has suggested that the protection of the home does not only apply to traditional houses, but will cover new forms of housing that have emerged since the 1950s, when the Convention was adopted. Among others, the right to home is not limited to residences which are lawfully established and may be invoked by a person living in a flat for which the lease is in the name of another tenant; the right to privacy may also be applied to social housing occupied by the applicant as a tenant, even though the right of occupation under domestic law has come to an end, or to the occupation of a flat for thirty-nine years without any legal basis. The protection offered under Article 8 ECHR is not limited to traditional residences and includes, for example, caravans and other non-fixed abodes, including cabins and bungalows occupying land, regardless of whether such occupation is lawful under domestic law; it may also cover second homes or holiday homes and even a legal person, such as a company, can invoke the right to 'home' when its business premises is entered.[14]

The right to data protection is no exception in this respect. As touched upon in the introduction, both legislators and courts have been prepared to widen the scope of the various definitions relevant to the right to data protection and have usually adopted a flexible approach when determining whether a certain data processing operation, technique or application would fall under a certain category or not. Not only has the scope of 'personal data' been broadened over time, the European Court of Human Rights has also accepted that under certain conditions, the secrecy of communication not only applies to the content of communication but also to communication data and metadata, because it is

> not persuaded that the acquisition of related communications data is necessarily less intrusive than the acquisition of content. For example, the content of an electronic communication might be encrypted and, even if it were decrypted, might not reveal anything of note about the sender or recipient. The related communications data, on the other hand, could reveal the identities and geographic location of the sender and recipient and the equipment through which the communication was transmitted. In bulk, the degree of intrusion is magnified, since the patterns that will emerge could be capable of painting an intimate picture of a person through the mapping of social networks, location tracking, Internet browsing tracking, mapping of communication patterns, and insight into who a person interacted with.[15]

Although the various legal categories are interpreted flexibly, they still guide and provide a basic framework for the reasoning of courts. Some of the most important legal differentiations are:

- *Personal data and non-personal data:* As explained, the distinction between personal data and non-personal data, even although the amount and type of data that are said to fall under the first category have grown exponentially over the last few decades, is determinative for the question of whether the General Data Protection Regulation applies. When non-personal data are processed, not only is there no protection for citizens but countries are prohibited from laying down restrictions and conditions for the processing and transfer of non-personal data. This has implications for two subsets of non-personal data:

    - *Identifying and anonymous data:* All anonymous data are non-personal data but not all non-personal data are anonymous data. Anonymous data are data that were personal data at one point but are no longer so, while non-personal data can also refer to data that never were personal data. Anonymisation is the process of stripping a dataset, however big or small, from all relevant identifiers. If the sentence 'Beppe Grillo is a dangerous politician' is changed to 'Mr. X. is a dangerous politician', where no other data are processed that could indirectly identify Grillo, the data are considered non-personal.

  - *Individual and aggregated data:* Aggregated data are non-personal data when the group of the category, the *n*, is large enough. If statistical analysis about 100,000 persons results in the profile '70% of the men who drive a yellow car are left-wing voters', such would not be considered personal data.

- *Personal data and sensitive personal data:* There is a special regime under the GDPR for 'sensitive data', which are 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. Because these data have a particularly close link to the individual, the GDPR specifies that processing them is in principle prohibited.[16] Although there are exceptions, these are more limited and restrictive than the grounds that can be used for legitimizing the processing of non-sensitive personal data.[17]

Content data and metadata: Though processing of data will only fall under the protection of private life if it significantly affects or harms a person, there is no threshold for the secrecy of communications. 'The content and form of the correspondence is irrelevant to the question of interference. [] There is no de minimis principle for interference to occur: opening one letter is enough.'[18] This means that even if a private email is opened that reads 'Honey, I'm having train delay and will arrive 30 minutes late', such would be considered an interference with the right to privacy under Article 8 of the European Convention on Human Rights. Although the ECtHR has stressed that there are situations in which metadata are collected in bulk and consequently fall under the scope of the right to privacy, such will only be the case when trails of metadata can be used to create detailed insights in the private life of a person. Consequently, most cases that concern the collection of metadata are not covered or are only marginally covered by Article 8 ECHR.[19] Likewise, metadata can fall under the protective scope of EU data protection law, though when they do, they enjoy different levels of protection.[20]

In short, legal regimes differentiate between types of data and relate to them different levels of protection. The processing of personal data is regulated, the processing of non-personal data is not; processing sensitive personal data is regulated more tightly than the processing of non-sensitive personal data; the bulk collection and use of metadata can fall under the scope of Article 8 ECHR, but only if it paints a clear picture of a person's private life, while the collection and use of content communication data will always be covered by the right to privacy; etc.

Such an approach to data regulation works well in a world where the nature of the data is relatively stable. This presupposition is challenged by Big Data processes. Big Data, for the purposes of this chapter, will be understood as data-driven processes that run through three phases.

1 Gathering: With respect to the volume of data, the basic philosophy of Big Data is ‘the more, the merrier’. The larger the data set, the richer the patterns and correlations that can be found and the more valuable the conclusions that can be drawn therefrom. Relying on smart computers and self-learning algorithms, artificial intelligence can learn from continuous data input and become ‘smarter’. Big Data can not only work on collected data, it can also produce new, inferred and probabilistic information. With respect to the variety of data sources, Big Data can be used to link an existing database to a database of another organization or to enrich it with information scraped from the internet. Because Big Data revolves around analysing large amounts of data and detecting general patterns and high-level correlations, the quality of specific data is said to become less and less important – quantity over quality. Because data gathering and storage is so cheap, data are often gathered without a predefined purpose; often, organisations determine only afterwards whether data represent any value to them and if so, to what use the data can be put.
2 Analysing: Once the data have been collected, they will be stored and analysed. The analysis of data is typically focussed on finding general characteristics, patterns and group profiles (meaning groups of people, of objects or of phenomena). General characteristics can be gained from data analytics – for example, how earthquakes typically evolve, from which indicators can predict an upcoming earthquake and determine which type of building is relatively unaffected by an earthquake. An important characteristic of Big Data is that the computer programs used for analysing data are typically based on statistics – statistical correlations are produced, not causal relations. These correlations typically involve probabilities. For example, an algorithm can predict that of the houses built with a concrete foundation, 70% will remain intact after an earthquake, while of the houses without a concrete foundation, this only holds true for 35%; or that people who place felt pads under the legs of their chairs and tables on average repay their loans more often than people who do not use felt pads. This also brings another point to the fore, namely that with Big Data, information about one aspect of life can be used for predictions about other aspects that are normally conceived as unrelated or belonging to a different domain of life. It may appear, for example, that the colour of a person’s couch has a predictive value for her future health, that the music taste of a person’s friends on Facebook says something about her sexual orientation, or that the name of a person’s cat has a predictive value for her career path.
3 Usage: The correlations gained through data analytics can be used at a general level. For example, when policy choices are based on the

> prediction that in 20 years' time, the majority of the population will be obese; they can be used to make predictions about groups of people, events or objects, such as bridges, immigrants or men with red cars and big houses; and they can be applied to specific, individual cases, projecting the general profile on a specific case. Well known data-driven applications include mass surveillance, predictive policing, smart cities, living labs, social credit scoring and personalised advertisements.

What is important to underline for the purpose of this chapter is that the nature of the data in Big Data processes is not stable, but highly volatile. A dataset that contains ordinary personal data may be linked and enriched with another dataset and become a set that contains sensitive data; the data may then be aggregated or stripped from their identifiers and become non-personal data; subsequently, the data may be deanonymised or integrated in another dataset containing personal data. The subsequent steps may happen in a split second. For example, when discussing the groups and categories in Big Data processes, it has been suggested that

> in the big data era, groups are increasingly fluid, not only through their changing membership, but also because of the changing criteria for the group itself. A group, the criteria for grouping people and the membership of a group might change in a split second. The purpose for which the group is designed may also change from day to day to adapt to new insights gained from data analytics, and groups may be formed and dissolved through the push of a button.[21]

This means that it becomes increasingly difficult to work with and uphold the various categories used in the law. The question is not only what falls under the definition of 'personal data', 'metadata', 'anonymous data' or 'sensitive personal data'; the point is that even though it might theoretically be possible to determine the status of a datapoint at every specific moment in time, this would be undoable in practical terms and defeat its purpose in legal terms, because applying a level of protection to a dataset at a specific moment in time is fruitless if its status is changed within a split second, potentially even a number of times. To draw a comparison: The question is not whether a caravan can, under specific conditions, also be considered a home deserving protection under the right to privacy, Article 8 of the European Convention on Human Rights. The question is whether it makes sense to work with a concept of a home, as distinguished from non-homes, when a specific building is to be considered a home at one moment in time, a business premises the next second, then a sex shop, a hospital next, a home again the following second, and so forth.

3. The status of data is unclear

The definitions used in the current legal framework have an element of indeterminateness. For example, the definition of personal data contains reference to 'identifiable information', which means that data that at this moment in time do not identify anyone, but may do so in the future, will be considered personal data nevertheless when the identification would cost relatively little effort. The other way around, in order to answer the question of whether data should be considered anonymous, an account should be kept of the efforts and investments needed to de-anonymise the data. As the former Working Party 29 explained,

> 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.[22]

Big Data has a number of important consequences for this constellation. First, not only is it possible to change the status of data and datasets within a split second due to the massive computational power and artificial intelligence, undoing these changes is also increasingly easy and cheap. To give an example, in 2010, Paul Ohm conducted a study on anonymisation techniques and discussed three cases in which organisations had made public databases which had been stripped from all identifiers; in each case, third parties, such as academics and journalists, where able to re-identify the people in that database by combining those data with other data. 'Each researcher combined two sets of data – each of which provided partial answers to the question "who does this data describe?" – and discovered that the combined data answered (or nearly answered) the question.'[23] Because of the increased technological powers to harvest indirect identifiable data and to combine existing databases with other open data sources, Ohm was convinced that in order to truly make a dataset anonymous, it has to be stripped from almost all data, hence arriving at the conclusion: 'Data can be either useful or perfectly anonymous but never both.'[24]

[EZ-Edit Graphic 15032-4199–006_Figure_001 here]

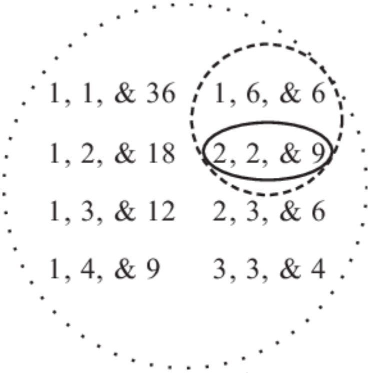

Figure 6.1 Census taker puzzle as an example of composition effects.

Building on this line of argument, Aron Fluitt and colleagues have recently suggested that more attention should be paid to what they call the composition problem of privacy. They explain this problem in basic terms by referring to a classic math puzzle:

> A man opens his door to a census taker, who asks how many people reside at the address and their ages. The man explains that it is just him and his three daughters. Instead of providing his daughters' ages, the man tells the census taker, 'The product of my daughters' ages is 36, and the sum is 13.' He then dismisses the census taker, noting 'I have to get my oldest daughter to her piano lesson.' The census taker thanks the man and accurately records the daughters' ages in his notes. – How was the census taker able to deduce the daughters' ages from the information provided? Each piece of information – the product of the ages, the sum, and the existence of an oldest daughter – narrows down the possible age combinations and ultimately reveals the exact ages. Figure 6.1 illustrates with a dotted circle the possible combinations of the three daughters' ages with a product of 36. Of those, the dashed circle contains the possible age combinations with a sum of 13. The solid circle contains the only combination that also has an oldest child: 2, 2, and 9. Although the possible age combinations that satisfy each clue independently may seem overwhelmingly vast, together the three clues eliminate all but one set of possible age combinations.[25]

This means that two databases that do not contain personal data by themselves, may when combined. To build on this example: If governmental database A contains information about a small village in which three people lived owning a yellow Ferrari, the school's database B provides that there is one family that always drives their four children to school in a yellow Ferrari, and child protection database C provides that in that village, there is one family with four children, of which there may be signs of domestic violence.

Connecting those and other dots may give a very detailed and insightful picture.

> [R]esearchers in 2018 revealed that the underlying confidential data from the 2010 US Decennial Census could be reconstructed using only the statistical tables published by the Census Bureau. They demonstrated a type of attack, called a *database reconstruction attack*, that leveraged the large volumes of data from the published statistical tables in order to narrow down the possible values of individual-level records. The researchers were able to reconstruct the sex, age, race, ethnicity, and fine-grained geographic location (to the block-level) reported by Census respondents *exactly* for 46% of the US population. They also showed that, if they relaxed their conditions and allowed age to vary by up to only one year, these five pieces of information could be reconstructed for 71% of the population. Further, the researchers showed that the reconstructed records could be completely *re-identified* – meaning they were able to assign personally identifiable information to individual records – using commercial databases available at the time. They concluded that, with this attack, they could putatively re-identify 138 million people, and they confirmed that these re-identifications were accurate for 52 million people, or 17% of the US population. These findings are startling. The last time the Census Bureau performed such a simulated re-identification attack on census datasets, the re-identification rate was only 0.0038%. The 2018 test attack demonstrates that previous risk assessments underestimated the re-identification risk by a factor of at least 4,500![26]

With the push to create an open data environment, in which datasets are published and made available for re-use,[27] still other datasets are available upon request or purchase and a high number of born digital data are generated on public websites, discussion fora and social network sites, enriching, merging and combining existing datasets becomes increasingly easy. The fact that technologies are ever more potent and the costs for operating algorithms have dwindled, means that both data and data-driven technologies are democratised. As explained at the beginning of this section, the status of the data (e.g., non-personal or personal) is determined in part by its to-be nature, which is dependent on the investments likely to be made by parties that have access to the data. Given that in the open data environment, datasets are increasingly made public, shared or made available upon purchase, it will be increasingly difficult to determine who will likely gain access to a dataset and what that party will do with the data.

But two things may be tentatively stipulated. First, given the democratisation of technologies and the minimal investment needed, it is increasingly likely that whenever a database is shared, there will be some party or another that will combine those data with other data, enrich them with data scraped from the internet or merge them into an existing dataset. Thus, although there is no certainty, it is increasingly likely that if an anonymised dataset is published, there will be some party around the world that will de-anonymise it or combine the data with other data in order to create personal profiles; that when a set of personal data is shared, there will be some party that will use those data to create a dataset with sensitive personal data; etc. Second, there will be other parties that have access to those data but will not engage in those types of activities; parties that will not use the data, use them as they are made available or even de-identify a database containing personal data. Who will do what is unclear beforehand.

Consequently, the legal category the data belong to is no longer a quality of the data itself, but a product of an organisation's efforts and investments. This means that it is unbeknownst beforehand whether an organisation will invest time and energy to harvest a database, and thus what the legal status of the data is. What is known is that it is increasingly easy and affordable to do so and that consequently, the likelihood that non-personal data may become personal data, that personal data may become sensitive personal data, that metadata will be used to create content data and that anonymous data will be re-identified is increasingly high. Applying the current legal categories strictly might mean that indeed almost all data should be seen as personal data and potentially as sensitive personal data, as there will most likely always be parties that will invest enough time, energy and resources to enrich a database. In addition, because data are increasingly available, shared and made public, the same database may have multiple legal statuses at the same time.

To draw from the analogy of the protection of the home again, the difficulty is not only, as described in the previous section, that the status of a building can change in a split second from a home to an office building to a fitness club to a private sex club to a home again. In addition, when determining whether a building should deserve the protection of a home, its future use should be taken into account; and while it is unknown whether the building will be used in the future as a home is unclear, it is increasingly likely that it will, though by whom is uncertain. Furthermore, the same building may have multiple functions for multiple parties at the same time, being a home to some, a restaurant to others, etc.

4. The status of the data is insignificant

The core rationale behind having the various data categories in law is their link to the individual. In general, the more directly data or datasets are linked to an individual and the more sensitive the data are, the higher the level of protection provided. To give an example, one of the first legal instruments to introduce the category of sensitive personal data was the Council of Europe's 1981 Convention. This introduction was elucidated in the explanatory memorandum in the following way:

> 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. Categories of data which in all member States are considered to be especially sensitive are listed in this article.[28]

The previous sections explained why the idea that the sensitivity of data is a quality of the data is increasingly redundant, but they did not question the core rationale underlying the legal categories. They showed that it is increasingly difficult to determine whether a set of information should be considered to contain non-personal, personal or sensitive personal data and that non-personal data may be converted into sensitive data in a split second, meaning that non-personal data should be seen as potential to-be sensitive data. This section will explain why the logic behind the various categories in law is increasingly redundant in the age of Big Data.

To provide an example, metadata can be just as revealing as content data, not just because they can reveal the content[29] – such as when a person visits a website with a xxx domain extension or when a letter is sent to a national cancer institute – but also because they reveal other information that may be even more sensitive than content data. The type of videos a person watches on a pornographic website may reveal one thing; the fact that the person either visits such a site once a year or twice a day may reveal more. What a person says to her mother over the telephone may reveal one thing; the fact that a person either spends two hours a day over the telephone or calls her mother once a year may say more.

The same argument applies to the fact that ordinary personal data may be as or more revealing than sensitive personal data. This is true not only because ordinary data may be used as proxies for sensitive personal data – such as that accurate predictions of someone's sexual preference can be based on her or even her online friends' music taste – but also because personal data may paint a highly personal picture of a person's life. For example, the fact that a person spends about eighteen hours a day on online gaming; the fact that a person stays up until 02.00 binge-watching Netflix series but logs into her work account at 07.00 from work; the fact that a person has founded six successive companies that all went bankrupt within a year; etc. may be more sensitive to many than the fact that they go to church, had a broken leg a year ago or are a member of a certain political party.

Admittedly, these are arguments that apply independently of the Big Data context. In addition, one could argue that what is considered sensitive or not is highly subjective, so that specified categories of sensitive personal data can never cover every aspect that one person or another may find sensitive, or that if in the Big Data era there would be new categories or types of data that should be considered sensitive, these could be included in the definition.

Consequently, the previous does not mean per se that the idea of designing a category of sensitive personal data is redundant; it only means that the category of sensitive personal data should be broadened or altered.

Still, Big Data allows organizations to connect metadata trails to gain detailed information about a person's life and allow harvesting of so many non-sensitive data about a person that a very granular picture about a person's life emerges. Not surprisingly, increasingly, companies and governmental organisations rely on gathering metadata rather than content communication, both because processing these types of data is subject to less restrictive rules and regulations and because the analysis of these types of data often yields more valuable results than the analysis of content communication data, among others because fewer datapoints are needed and because the datapoints are less ambiguous. In order to have an algorithm analyse content communication data, the program should be relatively well apt to understand natural languages used within specific contexts. It is far easier to create a heat map of where people go, how long they stay in specific places and who else is there; or which sites they visit, on what items they click, how long they stay on a specific page, etc.

Reference can also be made to the legal categories of aggregated or non-personal data and personal data. Increasingly, data analytics programs operate on anonymised and aggregated data or data that never were personal data. Big Data runs, as the word suggests, on large databases and the general lessons and patterns that are drawn therefrom. The correlations and group profiles may have as relevant determinant personal identifiers but are often based on non-personal datapoints, such as zip codes. Obviously, when such categories are used to the disadvantage of specific individuals, one may argue that data profiles should be considered personal data again. The classic reference here is to *redlining*, in which banks' policy on giving out loans was based on zip code areas, and the policy was disadvantageous to people living in neighbourhoods with a large African-American community. When a specific person is denied a loan on the basis of such a profile, it could be argued that this involves processing personal data.

Still, under such an approach, it is possible to design and make policies that affect groups of people on the basis of general information that were never personal data and may not have an effect on specific individuals but on large groups of society, or everyone living in society. For example, suppose an algorithm produces the result that one of the most effective ways to combat nighttime violence in a city is to spray a tangerine scent between 22.00–04.00 in nightlife areas because this makes people less aggressive. No personal data is processed, though such policies may have a high impact on people's lives.

In addition, because data protection regimes rely on the connection of the data to individuals and individual interests, two parts of the Big Data process are left unregulated. The gathering of non-personal or aggregated data is not regulated and the analysis of data, when correlations are found and group profiles are made, is left unregulated because this phase by definition revolves around analysing aggregated data. This holds true for the human rights regime in general. Referring to the example of redlining, the core of the problem is not that a particular black person is disadvantaged by the policy of the bank, but that the algorithm, the data or both are biased so that discriminatory policy emerges. Working with a biased dataset or a biased algorithm is currently not prohibited or sanctioned because analysing biased data or using biased algorithms as such does not harm any specific individual, and analysing that biased dataset with a biased algorithm is not regulated because of the focus on individual interests.

To refer to the metaphor of the home yet again, the reason for giving the home a special status was that within the private sphere, private matters were discussed, intimate actions took place and personal items were stored. If we are moving towards a world in which intimate actions take place irrespective of the physical domain, in which private discussions take place on open fora and in which personal items are stored in the cloud, then the question is whether the rationale behind the distinction between the private and the public domain is still valid. The same holds true for the categories in data protection law. If processing metadata can be just as or even more revealing than processing content data; if non-sensitive personal data can be put together in a way that gives a highly intimate picture of a person's life; if non-personal data can be used in ways that have far greater impact on the lives of ordinary citizens than the processing of sensitive personal data; then the question is whether the underlying rationale for the categorisations should be upheld.

5. Analysis

If these arguments hold true, two conclusions could be drawn. First, basing the level of regulatory protection on the status and nature of data is not the best way forward. Second, given the fact that non-personal data may be changed to sensitive data in a split second and that processing non-personal data can have a bigger impact on persons' lives than the processing of sensitive personal data, as long as the legal regulation *is* based on the status of data, it should provide for a basic framework for the protection of citizens' interests vis-a-vis the processing of non-personal data. This conclusion contrasts sharply with the approach taken by the European Union in 2018, when it adopted a Regulation on the transfer of non-personal data, which only aims at stimulating cross-border data processing, without providing any form of protection to citizens. Article 1 of that Regulation specifies: 'This Regulation aims to ensure the free flow of data other than personal data within the Union by laying down rules relating to data localisation requirements, the availability of data to competent authorities and the porting of data for professional users.'[30] The material provisions of the Regulation do not aim at restricting or laying down conditions for the processing or transfer of non-personal data but in contrast, prohibit any type of restriction or limitation in national laws on the availability, transfer and processing of non-personal data.

Given the previous, the EU might want to amend its approach and also provide protection to the interests of citizens when non-personal data are processed. The principles contained in the General Data Protection Regulation could serve as a source of inspiration. Although its material scope is determined by the identifiability of the data, many of the principles themselves do not so much aim to protect individual interests of specific data subjects but lay down general duties of care and standards for good data governance by data controllers, and can hence be transposed easily to the processing non-personal data.

For example, if an organization collects more non-personal data than it needs for its specified purpose, given that these data may be converted into sensitive personal data and that even the use of non-personal data can have a high impact on the lives of citizens, a data minimization principle could be applied to processing non-personal data all the same. Having a specific purpose for gathering non-personal or aggregated data and limiting the use of the data to that specific purpose seems a basic requirement in the age of Big Data. Given that increasingly, decisions are made on the basis of non-personal data and aggregated datasets are used to design policies, it seems vital to ensure that those aggregated data are correct, complete and up to date. In addition, given the fact that having and processing non-personal and aggregated data potentially provides organizations with just as much power as processing personal data, requirements to ensure transparency seem vital. In addition, as the impact of data processing operations based on non-personal data can be significant, an impact assessment also taking into account broader and societal interests may be regarded as quintessential in the age of Big Data, which also holds true for the requirement to appoint a Data Protection Officer. An obligation to ensure that the non-personal data are processed safely and securely, taking adequate technical and organizational security measures, having a data protection policy and embedding those principles in the technical infrastructure of an organization by design or default could help organizations abide by a regulation that would regulate processing non-personal data. Finally, like the current General Data Protection Regulation does, a Regulation on the processing of non-personal data should contain a rule specifying that transferring non-personal data to other jurisdictions should be prohibited, unless similar rules are applied to the processing of non-personal data.

Such a Regulation could be applicable to metadata and content data, sensitive data and insensitive data, anonymous and statistical data alike to the extent that they are not covered by the GDPR. The introduction of the Regulation on the Processing of Non-Personal Data would ensure that no data processing activity is left unregulated. It would also disallow Big Data organisations to circumvent data protection rules by temporarily aggregating data or stripping a dataset from identifying information.

A proposal for a regulation on the processing of non-personal data

Article 1 Applicability

This regulation applies to any natural or legal person that:

1 Processes non-personal data; and
2 Has an establishment in the European Union

Article 2 Definitions

For the purposes of this Regulation:

1 'non-personal data' means any information not relating to an identified or identifiable natural person;
2 'processing' means any operation or set of operations which is performed on non-personal data or on sets of non-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;
3 'data breach' means a breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, non-personal data transmitted, stored or otherwise processed

Article 3 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 non-personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures ('integrity and confidentiality')

Article 4 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')

Article 6 Transfers

The transfer of non-personal data to natural or legal persons outside the European Union is prohibited unless the person or organisation receiving the data signs a legally enforceable agreement in which that natural or legal person commits to upholding all principles contained in this Regulation.

Article 7 Enforcement

The tasks, powers and competences of the national supervisory authority and European Data Protection Board, as specified by the General Data Protection Regulation, shall also apply to the processing of non-personal data and the respect for the principles contained in this Regulation.

Notes

[1] International Bill of Human Rights: A Universal Declaration of Human Rights. <www.un.org/en/ga/search/view_doc.asp?symbol=A/RES/217(III)>.

[2] International Covenant on Civil and Political Rights. Adopted and opened for signature, ratification and accession by General Assembly resolution 2200A (XXI) of 16 December 1966. entry into force 23 March 1976, in accordance with Article 49. <www.ohchr.org/en/professionalinterest/pages/ccpr.aspx>.

[3] Convention for the Protection of Human Rights and Fundamental Freedoms Rome, 4.XI.1950. <www.echr.coe.int/Documents/Convention_ENG.pdf>.

[4] ECmHR, Campion v. France, application no. 25547/94, 06/09/1995. Unofficial Translation: In order to determine in similar cases the extent of the guarantee granted by article 8 (art. 8) against interference by public authorities, the Commission examines whether the taking of photographs constitutes an intrusion into the private sphere of an individual (for example when these were taken at her home), if the photographs refer to private or public events, and if they are intended to be used for a limited purpose or they are likely to fall within the public's knowledge. In the present case, the Commission notes that the photograph for which the applicant complains was taken on the public highway, when he was traveling by car, for the purpose of proof and identification. Nothing indicates that the photograph has been brought to the attention of the public or used for any purpose other than the prosecution of which the applicant has been the subject. Applying the criteria set out earlier, the Commission comes to the conclusion that there has been no interference with the privacy of the applicant.

[5] Secretary's Advisory Committee on Automated Personal Data Systems, Records, Computers and the Rights of Citizens (1973).

[6] U. Dammann, O. Mallmann & S. Simitis (eds), *Data Protection Legislation: An International Documentation* (Engl) (Frankfurt am Main: Metzner, 1977). F. W. Hondius, *Emerging Data Protection in Europe* (Amsterdam: North-Holland, 1975). H. Burkert, *Freedom of Information and Data Protection* (Bonn: Gesellschaft für Mathematik und Datenverarbeitung, 1983).

[7] 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). <https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680502830>. 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). <https://rm.coe.int/16804d1c51>.

[8] Charter of Fundamental Rights of the European Union (2000/C 364/01). <www.europarl.europa.eu/charter/pdf/text_en.pdf>.

[9] Article 1 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. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31995L0046>.

[10] 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). <https://eur-lex.europa.eu/eli/reg/2016/679/oj>.

[11] Article 2 sub a Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, Strasbourg, 28 January 1981. <https://rm.coe.int/1680078b37>.

[12] See for a discussion: B. van der Sloot, *Privacy as Virtue* (Cambridge: Intersentia, 2017).

[13] Article 1 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. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32018R1807>.

[14] Guide on Article 8 of the European Convention on Human Rights Right to respect for private and family life, home and correspondence Updated on 30 April 2019. <www.echr.coe.int/Documents/Guide_Art_8_ENG.pdf>.

[15] ECtHR, Big Brother Watch and Others v. The United Kingdom, application nos. 58170/13, 62322/14 and 24960/15, 13 September 2018, § 356.

[16] Article 9 GDPR.

[17] Article 6 GDPR.

[18] Guide on Article 8 of the European Convention on Human Rights Right to respect for private and family life, home and correspondence Updated on 30 April 2019. <www.echr.coe.int/Documents/Guide_Art_8_ENG.pdf>.

[19] Another source of inspiration could be the EU's e-Privacy Directive, which differentiates between traffic data, that are defined as any data processed for the purpose of the conveyance of a communication on an electronic communications network or for the billing thereof, location data, meaning any data indicating the geographic position of the terminal equipment of a user of a publicly available electronic communications service, location data other than traffic data, etc. Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications).

[20] Joined cases C-203/15 and C-698/15 Tele2/Watson [2016] ECLI:EU:C:2016:970, para 99. For a full analysis of this CJEU judgment, see Will R. Mbioh, 'Post-och

Telestyrelsen and Watson and the Investigatory Powers Act 2016' (2017) 3 (2) *EDPL* 273–282.

[21] L. Taylor, L. Floridi & B. van der Sloot (eds), *Group Privacy* (Dordrecht: Springer, 2017), p. 284.

[22] Article 29 Data Protection Working Party, Opinion 4/2007 on the concept of personal data, 01248/07/EN WP 136, 20 June 2017. <https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf>. See also: Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques, 0829/14/EN, WP216, 10 April 2014. <https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf>.

[23] P. Ohm, 'Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization' (2010) 57 *UCLA Law Review* 1701, p. 1723.

[24] Ibid, p. 1704.

[25] A. Fluitt et al., 'Data Protection's Composition Problem', (2019) 5 (3) *European Data Protection Law Review*.

[26] Ibid.

[27] See inter alia: Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information. <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:345:0090:0096:en:PDF>. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32013L0037&from=FR>.

[28] Explanatory Report to the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data Strasbourg, 28 January 1981. <https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016800ca434>.

[29] B. Greschbach, 'The Devil is in the Metadata – New Privacy Challenges in Decentralised Online Social Networks'. <www.nada.kth.se/~gkreitz/metadata/sesocMetaPrivacy.pdf>.

[30] Article 1 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. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32018R1807>.