FACULTY OF LAW
Stockholm University

BIG DATA
- How We Can Utilize its Benefits

Alice Castler

Thesis in Legal Informatics, 30 HE credits
Examiner:
Stockholm, Autumn/Spring term 2016/2017

Stockholm
University

ABSTRACT

Big data analytics offers tremendous value to society and are the new driving force for innovation, productivity, efficiency and growth. However, big data containing personal information implies privacy concerns. Within the European Union, the member states have addressed these privacy concerns by adopting rather extensive rules on the protection of personal data. In an environment where everything is recorded, stored, analyzed and shared, far-reaching data protection legislation is much needed. However, the extensive limitations on the processing of personal data, which are entrenched both in the existing Data Protection Directive and the upcoming General Data Protection Regulation, prevent society from fully utilizing the benefits of big data. It is nearly impossible to find correlations that lead to groundbreaking discoveries and at the same time adhere to these limitations. It can hence be questioned whether strong privacy protection and beneficial uses of data at all can coexist in today's society. Both the Data Protection Directive and the General Data Protection Regulation are intended to strike an adequate balance between these two concepts by recognizing a possibility to render personal data non-personal through the method of anonymization. However, during the last years, computer scientists have shown that anonymized data often can be reidentified. The revelation of this reidentification risk has made it more difficult than ever to reach the requisite level of anonymization and thus to avoid application of the legislation. This entails that more data than originally intended is now falling under the scope of the legislation, which leaves less room for utility. Hence, the balance between privacy and utility in European data protection legislation has been disrupted. Clearly, something needs to be done in order to restore that balance. This thesis encourages the establishment of contextually sensitive standards, which state how certain data should be anonymized and what other safety measures need to be taken in order to avoid application of the legislation. This thesis further suggests an introduction of guarantees, meaning that a dataset is guaranteed to fall outside the scope of the legislation if all requirements in a standard targeting the relevant dataset are fulfilled. By establishing such guarantees the current imbalance between privacy and utility in European data protection legislation can be decreased and society can retain the possibility to utilize the benefits of big data.

Keywords: *Big Data, EU Data Protection Legislation, Anonymization, Reidentification, Privacy, Utility*

LIST OF ABBREVIATIONS

A29WP	Article 29 Data Protection Working Party
CFREU	Charter of Fundamental Rights of the European Union
CJEU	Court of Justice of the European Union
DPD	Data Protection Directive
ECHR	European Convention of Human Rights
EDPB	European Data Protection Board
EU	European Union
GDPR	General Data Protection Regulation
ICO	Information Commissioner's Office
TFEU	Treaty on the Functioning of the European Union

TABLE OF CONTENT

1. INTRODUCTION

Twenty-seven years ago, the first web server and web browser was invented by Tim Berners-Lee.[1] Since then the amount of information produced has exploded. According to the latest statistics, we create 2.5 quintillion bytes of data every day and all the data existing in the world today has been created in the last two years.[2] This rapid growth is expected to continue and the statistics will probably show even higher numbers in a few years. The data comes from everywhere – from sensors in mobile phones and cars, from online transactions and credit card purchases, from search queries on web search engines, from emails, clickstreams and logs, but also from health records, electric grids, roads, bridges and global positioning satellites.[3] In other words, everything we do and everything that happens in the world is recorded. Data has grown to an enormous volume, contains extreme variations and are produced so fast that a new term has been established. The term referred to is *Big Data.* Different actors in the society have realized that data with these characteristics offers tremendous opportunities. By analyzing such data new discoveries and improvements have been made that no one thought were possible before. Gary King, director of Harvard's Institute for Quantitative Social Science, calls it a revolution.[4] So do Viktor Mayer-Schönberger and Kenneth Cukier, who have written the book *Big Data: A Revolution That Will Transform How We Live, Work and Think.* Data has become the raw material of production and a new source for immense economic and social value.[5] Data has even been declared a new class of economic asset, like currency or gold.[6] Hence, it can be concluded that data, and especially big data, can be extremely valuable and the future development of our societies rely, to a great extent, upon the analysis of such data. However, big data do not only imply benefits but also challenges. The fact that nearly every move we make is recorded and analyzed raises privacy concerns. To tackle the increasing privacy risks in society, data protection legislation has been adopted around the world. Within the European Union (EU) a directive on the protection of personal data was

[1] Rhiannon Williams, *Web Browsers: A Brief History*, The Telegraph, 2 May 2015, http://www.telegraph.co.uk/technology/microsoft/11577364/Web-browsers-a-brief-history.html, accessed 12 December 2016.
[2] IBM, *Bringing Big Data to the Enterprise: What is Big Data,* https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html, accessed 12 December 2016.
[3] Omar Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics,* 11 Nw. J. Tech. & Intell. Prop. 239, 2013, p. 240.
[4] Steve Lohr, *The Age of Big Data,* The New York Times, 11 February 2012, http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all, accessed 12 December 2016.
[5] Tene & Polonetsky (2013), *Big Data for All: Privacy and User Control in the Age of Analytics*, supra note 3, p. 239.
[6] Lohr, supra note 4. See also Paul M. Schwartz, *Property, Privacy, and Personal Data*, 117 Harv. L. Rev. 2055, 2004, p. 2056.

adopted already in 1995 (hereinafter called the DPD).[7] The directive will be replaced by a regulation called the General Data Protection Regulation (hereinafter called the GDPR), which comes into force on May 25, 2018.[8] Both the DPD and the GDPR limit the ways in which personal data can be processed by imposing extensive requirements on actors handling such information.[9] This implies that both the current and the future legislation deprive the society of the opportunity to fully utilize the benefits of big data in favor of protecting our privacy. Striking the right balance between beneficial use of big data and privacy risks has been called "the biggest public policy challenge of our time".[10] The question is whether privacy and big data really can coexist in our increasingly complicated world. This thesis examines and proposes new, innovative solutions to this challenge.

1.1 The Aim of the Thesis and Research Questions

As briefly touched upon above both the existing and the future data protection legislation within the EU imposes limitations on how personal data can be processed. The characteristics of big data analytics makes is very difficult, if not impossible, to adhere to these limitations. Hence, actors have strived to avoid falling under the scope of the law in order to be able to reap the benefits of big data. Until approximately a decade ago the application of the law could easily be avoided by anonymizing personal data. However, computer scientists have proven, in several cases, that anonymized data can be reidentified. This revelation raises the question whether we should continue to rely on anonymization for the purpose of enabling the society to fully take advantage of the opportunities with big data. Hence, the aim of this thesis is to examine whether existing anonymization techniques still are sufficient for rendering personal data non-personal and thus for exempting data from the scope of the legislation. If all anonymization techniques are proven to be insufficient in this regard, the aim is also to examine if there are any alternative solutions to the challenge of enabling society to utilize the benefits of big data while protecting every one's right to privacy. At its core, the questions being examined are whether EU data protection legislation strikes an adequate balance between privacy and utility[11], and if not, how that balance can be restored.

[7] 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.
[8] Regulation 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.
[9] See especially article 6 and 7 in the DPD and article 5 and 6 in the GDPR.
[10] Omer Tene & Jules Polonetsky, *Privacy and Big Data: Making Ends Meet,* 66 Stan. L. Rev. 25, 2013, p. 26.
[11] The word 'utility' is used throughout this work to refer to beneficial uses of data, especially of big data.

To be able to conduct this examination one must first fully understand the concept of big data and its benefits in greater detail, but also its challenges, which include the difficultness of harnessing the benefits of big data while complying with the data protection legislation. It is further necessary to present the basic ideas of anonymization, in order to understand how this method can be used to unlock the benefits of big data. However, it is also necessary to analyze each anonymization technique in greater detail, to be able to determine whether any of them are sufficient for exempting data from the legislation, or if alternative solutions need to be developed. The aim of the thesis can thus be expressed in the following research questions:

- What is big data?
- What are the benefits and the challenges of big data?
- What is anonymization and how does it work?
- Is there any anonymization technique that is sufficient for enabling the society to fully utilize the benefits of big data?
- If not, are there any alternative solutions?

1.2 Delimitations

Both personal data (i.e. all information that can be traced back to a living person) and non-personal data (e.g. climate data) can be used for big data analytics. However, the problem addressed by the author does not appear in case the data is non-personal in nature. Hence, this thesis will only focus on data that is originally personal data. Moreover, only the data protection legislation within the EU will be taken into consideration and not any other legal framework in the world. However, different conclusions made in the following may apply in other legal systems as well, since the basic rules on data protection often are rather similar. Furthermore, only a few provisions in the directive and in the regulation will be touched upon, in particular article 2, 6 and 7 in the directive and article 4, 5, 6, 83 and 99 in the regulation. Hence, there will be no complete presentation of the EU data protection legislation. The reason for this is that the problem addressed by the author does not directly concern any other provisions than those mentioned above. Lastly, anonymization techniques commonly discussed in legal and computer science literature will be evaluated below, but this thesis shall, however, not be seen as an exhaustive presentation of all anonymization techniques ever developed.

1.3 Method and Material

At its core, this thesis is composed using the method of legal dogmatics insofar as the positive law related to the research subject and its systematic order is presented and interpreted through an in-depth analyze of traditional legal sources (i.e. legislative acts, court cases and legal doctrine). The method of legal dogmatics is used in this work to present the content of law in an orderly manner with the aim of identifying flaws in the law. Hence, this thesis goes beyond the purely descriptive plane by identifying and analyzing problems in the law from a big data analytics perspective. In addition, this thesis discusses and proposes solutions to the identified problems. Thus, in this regard a more analytic approach is taken. Accordingly, the method applied in this work is nuanced from traditional legal dogmatics insofar as the positive law is being criticized and arguments are brought forward regarding what the law should be.

Moreover, the arguments presented are not only based on traditional legal sources, but instead there is dependence upon materials gathered in the domain of computer science as well.[12] To be able to answer the central research question regarding whether EU data protection legislation strikes an adequate balance between privacy and utility, one must first examine if any existing anonymization technique still is sufficient for exempting data from the scope of the legislation. This examination requires a rather comprehensive understanding of the technical details of different anonymization techniques, which is why materials derived from computer science are a necessary element in this thesis. Such materials are mainly used in Chapter 3 covering anonymization, which is written more from a computer-science perspective and thus adopts a slightly different terminology than the other chapters. The technical aspects presented in Section 3.1-3.2 are connected to a comprehensive legal analysis in Section 3.3 and Chapter 4-5. Accordingly, computer science is used as a "supporting discipline" in forming the legal arguments brought forward herein.[13]

To summarize, this thesis is composed using a nuanced version of legal dogmatics, which can be referred to as analytical legal method.[14] This method allows for a critical examination of the positive law as well as *de lege ferenda* proposals based on a broad selection of materials derived from not only the legal domain but also from other scientific disciplines.

---

[12] See Liane Colonna, *Legal Implications of Data Mining: Assessing the European Union's Data Protection Principles in the Light of the United States Government's National Intelligence Data Mining Practices*, Ragulka förlag, 2016, p. 37.
[13] See Bart Van Klink & Sanne Taekema, *On the Border: Limits and Possibilities of Interdisciplinary Research*, in *Law and Method*, Tubingen, Mohr Siebeck, 2011, p. 11.
[14] Claes Sandgren, *Rättsvetenskap för uppsatsförfattare: ämne, material, metod och argumentation*, Norstedts Juridik, 2015, pp. 45-47.

As noted above, a vast selection of sources has been used to compose this thesis. First of all, legislative acts of the EU, including both primary (i.e. the Charter of Fundamental Rights of the EU (CFREU) and the Treaty on the Functioning of the EU (TFEU)) and secondary (i.e. the DPD and the GDPR) legislation, have been used to determine the current state of law. To the extent possible jurisprudence of the Court of Justice of the European Union (CJEU) has been used for interpreting the legislation. In addition, opinions of the Article 29 Working Party (A29WP) have been employed to understand how the law applies.[15] Reports, surveys and comments from different public bodies and institutes are also referred to throughout the work. Furthermore, the research relies to a great extent upon the analysis of authoritative literature and journal articles written by leading scholars in the field of data protection.

Although this thesis is primarily based on legal sources, certain parts, especially Chapter 3, rely to some extent on materials derived from the domain of computer science. Such materials include literature and journal articles with authoritative value, but also articles from newspapers and magazines, different websites and dictionaries. As touched upon above, the interdisciplinary element in this thesis requires and allows a broader selection of materials than from the traditional legal sources.

1.4 Disposition

The thesis consists of five chapters, which disposition follows the research questions listed in Section 1.2. Chapter 1 comprises this introduction. Chapter 2 outlines what big data is, its benefits, how it threatens privacy and why it is particularly difficult to comply with the EU data protection legislation when conducting big data analytics. Chapter 3 first examines why anonymization in theory is a good strategy for reaping the benefits of big data, and secondly evaluates whether any anonymization technique actually is sufficient for that purpose. In Chapter 4 a few alternative solutions to the problem of harvesting the benefits of big data while complying with the data protection legislation are discussed. Finally, in Chapter 5 the author presents some concluding remarks.

[15] The A29WP is an expert organ, which task is to contribute to a uniform application of the DPD within the EU. For that purpose, the A29WP gives opinions on complicated legal issues. Even though these opinions are not binding, they provide useful guidance on how to interpret the law. See article 29 and 30 in the DPD.

2. BIG DATA

Big data is truly an intriguing phenomenon[16] and it is currently a major topic of discussion. Big data is being discussed in our newspapers, in scientific articles across a number of fields and within governments from several aspects. This single phenomenon has already given birth to a number of literary works. The attention big data has drawn in recent years is enormous and it has become clear that it does affect the world to a great extent. Although we stumble across the term *big data* quite frequently in our daily lives, one, at least the uninitiated, may find it hard to grasp. What does big data really mean? In fact, there is no given definition that allows us to clearly distinguish any particular instance of processing as big data or not. However, there are a few factors that particularly characterizes big data and which are often used to describe the phenomenon. These factors will be presented below with the aim to bring some clarity to the term big data, which can be found to be somewhat ambiguous. After the term has been described in greater detail, the value of big data will be analyzed, which will be followed by a discussion regarding the difficulties with big data analytics.

2.1 What is Big Data?

Even though there is no single definition established of the term big data, several definitions do exist. A commonly cited definition is one from the Gartner IT glossary, which defines big data as "high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making".[17] The definition provides that big data has three main characteristics, which are often called the 'three Vs' – *volume, variety* and *velocity*.[18] Moreover, the definition also points out that this type of information assets requires special analyzing methods. Hence, the three key characteristics as well as the method used for processing big data will be examined below.

2.1.1 Volume

First of all, big data refers to enormous amounts of data. The data contained in these massive datasets can be meta-data from internet searches, credit and debit card purchases, social media postings, mobile phone location data, or data from sensors in cars and other electronic devic-

[16] In fact, big data has been described as a phenomenon rather than a technology. See for example Leslie Wiggins, *If Big Data and Analytics Exist in a Silo, Does the Outcome Matter?*, IBM Big Data and Analytics Hub, 25 February 2014, http://www.ibmbigdatahub.com/blog/if-big-data-and-analytics-exist-silo-doesoutcome-matter, accessed 18 November 2016.
[17] Gartner IT glossary, http://www.gartner.com/it-glossary/big-data, accessed 18 November 2016.
[18] Information Commissioner's Office (ICO), *Big Data and Data Protection*, 2014, p. 6, available at www.ico.org.uk.

es.[19] What is significant for these types of datasets is that they cannot be analyzed using so called 'traditional methods', such as Excel spreadsheets or relational databases.[20] They are simply too big. To be able to understand how large amounts of data one is referring to when talking about big data, it is necessary to take a look at the development in recent years. The amount of data in the world has increased tremendously over the last years. In fact, 90 % of the data in the world today has been created only during the last two years.[21] The Boston Consulting Group estimated a total growth of 2.5 exabytes, which equals to 2.5 billion gigabytes, per day by the year of 2013.[22] How much data that is being produced and stored today, year 2017, is rather unclear, but most likely has it increased since year 2013 and a further rapid growth is expected. One might wonder what enabled this large expansion. The main reason is probably that it has become a lot easier to hold very large datasets, even for small private actors. The availability of data storage has increased, due to the emergence of cloud-based services, and the cost of storage has decreased.[23] Another aspect that has been very important for the usage of big data, is that new tools have been developed that can analyze such massive datasets.[24] Due to these factors, we are now able to utilize data in such a revolutionary way that no one could ever have dreamt of before.

2.1.2 Variety

Besides the fact that big data is characterized by the vast amounts of data being collected, it is also characterized by the great variety of information contained in one dataset. Data is often collected from a number of sources, which is then complied in one dataset.[25] The data can both be structured, for example in tables with defined fields, or unstructured in a rather text-heavy document.[26] For instance, it is rather common that businesses obtain unstructured information from social media sources regarding what their customers think about their products, which is then compared with sale statistics held by the company in structured form. The possibility to combine different types of information from different sources can be extremely valuable for a company seeking to develop and improve its products, but from an IT-

[19] Ibid.
[20] Ibid., p. 7.
[21] IBM, supra note 2.
[22] Robert Souza, Rob Trollinger, Cornelius Kaestner, David Potere & Jan Jamrich, *How to Get Started with Big Data,* BCG perspectives by the Boston Consulting Group, 29 May 2013, https://www.bcgperspectives.com/content/articles/it_strategy_retail_how_to_get_started_with_big_data/, accessed 19 November 2016.
[23] ICO (2014), supra note 18, pp. 6-7.
[24] Ibid., p. 7. Two commonly used tools are NoSQL and the open source software Hadoop.
[25] Ibid.
[26] Ibid.

perspective it can be rather complicated. However, technologies have been developed enabling the information to be analyzed and compared even if it is not contained in one single database structure.[27] Although the technological problems have been solved, the fact that data is being combined from several sources gives rise to certain privacy issues, which will be discussed later in section 2.3.

2.1.3 Velocity

These large amounts of different types of data are often produced, stored and analyzed at high speed.[28] Hence, the third key characteristic of big data is often said to be velocity. Big data analytics can both refer to analysis of data 'in motion', i.e. the data is being analyzed simultaneously as it is produced, and analysis of stored data, i.e. the analysis is carried out a certain time after the data has been produced.[29] The possibility to analyze data in real or near-real time is of great value to society. Governmental institutions and other organizations are able to analyze real time video feeds to identify security threats and companies can utilize real time information to improve their customer services by, for instance, sending a coupon to a customer standing in the cereal aisle based on the customer's past cereal purchases.[30]

Although volume, variety and velocity are the key elements in the concept of big data that are constantly emphasized, it is important to bear in mind that there is no fixed definition of big data.[31] A particular instance of data processing does not have to be significant in terms of all three components to classify as big data.[32] Moreover, there are several commentators who describe big data in terms of other criteria than those mentioned above, or present many more criteria.[33]

2.1.4 Analyzing Big Data

In order to fully understand the phenomenon big data, we need to take a deeper look into the analysis of such data. Big data, with its rather unique characteristics, demand, as provided by the definition in the Gartner IT glossary, cost-effective and innovative forms of processing.[34] Thus big data analysis differs significantly from traditional methods used for analyzing data. Before the event of big data, datasets were analyzed by constructing queries that were run

[27] Ibid.
[28] Richard Kemp, *Big Data and Data Protection,* White Paper, Kemp IT Law, 2014, p. 2.
[29] ICO (2014), supra note 18, p. 7.
[30] Souza et al., supra note 22.
[31] ICO (2014), supra note 18, p. 8.
[32] Ibid.
[33] See for example Kemp, supra note 28, pp. 2-3.
[34] Gartner IT Glossary, supra note 17.

against the dataset in order to derive answers to predefined questions.[35] Hence, this method required you to know beforehand what you were looking for. In big data analytics the opposite situation applies, you do not necessarily need to know what you are searching for. Big data is analyzed by running a very large number of algorithms against the data in order to find meaningful correlations and patterns between variables.[36] Unlike traditional forms for processing data, analytic methods employing algorithms, called data mining, does not require a hypothesis to commence the analysis.[37] Hence, the results of such analysis are very unpredictable and can reveal patterns and correlations that no one could have thought of before.[38] This can both be of great value and imply certain challenges, which will both be discussed below in Section 2.2 and 2.3.

2.2 The Benefits of Big Data

As briefly touched upon above big data can be of great value. The most intriguing aspect is that big data can be highly beneficial in every single area within our society. Both the public sector as well as the private sector can through big data analytics achieve important improvements, which in the end will benefit every one of us. This is the reason why big data has been described as a revolution that will transform the world.[39] The tremendous value of big data can be illustrated by the following few examples.

An often-cited example is Google Flu Trends. In 2009 a new virus was discovered, called H1N1, which contained elements of both the viruses that cause bird flue and swine flu. H1N1 spread quickly and many feared an outbreak of a terrible pandemic that could plausibly kill millions of people. An unease spread amongst people since no vaccine was available that could cure the disease. There was nothing left to do other than for the public health authorities to try to slow down the spread of the virus. However, in order to pursue this goal, the health authorities needed information regarding where the virus already had spread. In the United States, information was gathered by requesting doctors to report any new flu cases to the public health authority for disease control and prevention. However, by the time the information reached the authority it was already outdated, since people in most cases did not consult a doctor directly when the first symptoms occurred and reporting the information back to the

[35] ICO (2014), supra note 18, p. 8.
[36] Tal Z. Zarsky, *Desperately Seeking Solutions: Using Implementation-based Solutions for the Troubles of Information Privacy in the Age of Data Mining and the Internet Society*, 56 Maine Law Review 13, 2004, pp. 27-28.
[37] Ibid.
[38] Ibid.
[39] See for example Viktor Mayer-Schönberger & Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work and Think*, John Murray, 2013.

central authority took time. Hence, the public authorities seemed to be failing in taking control over the flu. The method for gathering the data was simply not sufficient. It lacked in terms of volume, variety as well as velocity. While the public authorities struggled with their outdated information, the Internet giant Google used big data analytics to get a more accurate picture of the pandemic that was emerging. Google could, by analyzing the vast amounts of search queries they received every day[40], predict the spread of the virus in the whole United States. By comparing the search queries with statistics on the spread of seasonal flu over the last years, Google, by employing different mathematical models, found correlations between their predictions and the official figures. Hence, unlike the public authorities, Google manage to tell where the flu had spread, not one or two weeks too late, but in near real time, and thus Google's research proved to be of greater value than the public authorities'. Google Flu Trends is an excellent example that shows how big data analytics can be used to gain information of great importance and thus be of significant value to society.[41]

Big data can make a huge difference, not only within the healthcare sector, but in various areas in the society. Big data is, for instance, used to foresee climate changes like the rise of the sea level and to help government agencies to detect fraud.[42] Furthermore, big data can be used in an extremely powerful way within the business sector and companies view this information as a corporate asset of significant value.[43] Big data is described as the "the next frontier for innovation, competition and productivity" and is currently transforming the global economy.[44] There are several ways in which big data creates opportunities and enables improvements within the business sphere. By analyzing large amounts of data from different sources, companies can discover needs of customers and clients that where unknown before.

[40] In fact, Google receives three to four billion search queries every day. For live Google search statistics go to http://www.internetlivestats.com/google-search-statistics/, accessed 24 November 2015.
[41] The example is taken from Mayer-Schönberger & Cukier, supra note 39, pp. 1-2. Although Google Flu Trends serve as an excellent example of the positive benefits of big data, the project has been hardly criticized. Several reports suggest that Google Flu Trends drastically overestimated the 2012 and 2013 flu season. See for example Declan Butler, *When Google Got Flu Wrong: US Outbreak Foxes a Leading Web-based Method for Tracking Seasonal Flu,* 494 Nature 155, Macmillan Publishers Limited, 2013. See also David Lazer & Ryan Kennedy, *What We Can Learn From the Epic Failure of Google Flu Trends*, Wired Science, 2015, available at https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/. Moreover, Paul Ohm argues that since the results of the project was only shared through very few channels, the purpose could not have been to save lives but merely to market Google and thus the violation on privacy cannot be justified. See Paul Ohm, *The Underwhelming Benefits of Big Data,* 161 U. Pa. L. Rev. 339, 2013, p. 342.
[42] Abu Bakar Munir, Siti Hajar Mohd Yasin & Firdaus Muhammad-Sukki, *Big Data: Big Challenges to Privacy and Data Protection*, 9 International Journal of Social, Education, Economics and Management Engineering 355, 2015, p. 355.
[43] Schwartz (2004), supra note 6.
[44] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh & Angela Hung Byers, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, Report – McKinsey Global Institute, 2011.

They can also discover how the needs changes and thus adjust their products or services accordingly.[45] Moreover, within the analyze it is possible to segmenting populations in order to fully customize the products and the services for current and future customers. As a result of the analysis it is even possible to support or replace human decision making with automated algorithms.[46] All these possible measures enable improvements of products and services for the benefit of both business actors and individuals. However, big data analytics does not only allow improvements but has already and will probably continue to give rise to completely new products, services and entire business models.[47]

A good example of how big data analytics can be used to create new services is the story behind the company Farecast, started by the computer scientist Oren Etzioni and later acquired by Microsoft. It all started with a flight from Seattle to Los Angeles, on which Etzioni compared the price of his own ticket with the price his co-passengers had paid for their tickets. Etzioni realized that he had paid a lot more than all the other passengers he asked, although the others had purchased their tickets much more recently than Etzioni had bought his. From this experience Etzioni created a predictive model that were able to determine whether a certain ticket price seen online was a good or a bad deal. When developing this model Etzioni used a sample of 12 000 price observations from a travel website over a 41-day period. However, when this little project evolved into a startup named Farecast much more data was needed. Etzioni obtained access to one of the industry's flight reservation databases and thereafter based the predictions on nearly 200 billion flight-price records. By the help of big data Etzioni could develop a service that armed people with information they could only have dreamt of having access to before and thus was a valuable support in deciding weather to buy or not to buy a certain plane ticket. To conclude, by having access to vast amounts of data and having the ability to analyze that data, Etzioni saved consumers a bundle while making a fairly good profit himself.[48] Hence, it is clear that big data increases innovation, competition and productivity to the benefit of private enterprises, consumers and the global economy as a whole.[49]

As touched upon above, big data is not only a powerful phenomenon for governments, companies and other organizations, but also for individuals. Today, almost every one of us has immediate access to enormous amounts of data from every corner of the world through our beloved smartphones. This means that people can make more well informed decisions,

[45] Munir, Yasin & Muhammad-Sukki, supra note 42, p. 356.
[46] See for example the case with the customer standing in the cereal aisle mentioned on page 13 above.
[47] Munir, Yasin & Muhammad-Sukki, supra note 42, p. 356.
[48] In fact, Microsoft bought Farecast for around 110 million US dollars.
[49] The example about Etzioni is taken from Mayer-Schönberger & Cukier, supra note 39, pp. 3-5.

can come up with more innovative ideas and also communicate these thoughts to the rest of the world. The trend we are witnessing with the constantly increase in data flourishing around in our society has also a great impact on basic human rights, such as the freedom of information and the freedom of expression.[50] These rights are automatically strengthened as the increase of data makes the world more transparent. Hence, the existence of big data not only benefits individuals as customers or clients in relation to other actors, but also as mere human beings.

To summarize, big data offers a lot of benefits to all of us living on this planet and the powerfulness of this phenomenon cannot be exaggerated. As evidenced by the examples presented above, big data really creates new opportunities within every area of the society, within medical research, national security, marketing and urban planning, just to mention a few.[51] However, its powerfulness also implies challenges and for reasons to be mentioned below there is a need to strike a balance between the interest to utilize the benefits of big data and other conflicting interests.

2.3 The Challenges with Big Data

Although big data offers tremendous opportunities, it also includes challenges. As a matter of fact, it can imply severe privacy concerns.[52] We are spreading information that can be traced back to us wherever we go and whatever we do. For instance, when we need access to a Wi-Fi on a public location, we type in our e-mail address and sometimes even where we are from, where we live and what gender and age we are. However, we do not only spread personal data by actively entering information in a digital forum, but also as we simply go about with our daily lives. Nearly every step we take is being registered nowadays – what we purchase, where we travel, what music we listen to, what movies and TV-series we watch, which websites we visit etc. In fact, all our online activity is being deeply scrutinized, which may reveal quite a lot about a specific person.[53] In addition, even more sensitive data is gathered about us, such as regarding our health and exact location.[54] All the data that we leave behind is collected, stored, and analyzed by different actors in the society. The data is even shared or sold to third parties finding the information interesting and useful. All these activities threaten our privacy in different ways.

[50] Munir, Yasin & Muhammad-Sukki, supra note 42.
[51] Tene & Polonetsky (2013), *Privacy and Big Data: Making Ends Meet*, supra note 10, p. 25.
[52] Ibid. See also Omer Tene & Jules Polonetsky, *Privacy in the Age of Big Data: A Time for Big Decisions,* 64 Stan. L. Rev. Online 63, 2012, p. 65.
[53] Tene & Polonetsky (2012), supra note 52.
[54] Ibid.

Privacy can be understood as encompassing a right to seclusion or to be left alone, a right to non-interference in decision-making and a right to control over one's personal information.[55] The fact that information about us is gathered and stored disrupts our right to seclusion. It can also create a feeling of being under constant surveillance, affecting which activities individuals engage in, and hence, can be seen as interfering with one's decisional privacy.[56] Moreover, the analysis of personal information may imply further privacy concerns. Looking at only one piece of information isolated may not reveal that much about an individual, but when combined with other pieces of data as in big data analytics, sensitive information can be inferred about an individual.[57] With all the data available in society today, it is even possible to draw up a pattern of someone's everyday life.[58] This certainly disrupts the right to be left alone and can cause serious harm to individuals if sensitive information is revealed in an unwanted context.[59] Furthermore, the sharing of personal data may interfere with the right of having control over one's personal information.[60] Hence, clearly big data analytics, although highly beneficial, imply severe privacy concerns.[61]

In today's society, where the majority of actions taken in our daily lives are being registered and analyzed to scrutiny, it can be questioned whether there is room for privacy at all. However, privacy has for many years been considered a fundamental human right within the EU and thus is strongly protected. In fact, every nation within the EU is bound by the European Convention on Human Rights (ECHR) to protect the basic right to respect for private and

[55] Antoinette Rouvroy & Yves Poullet, *The Right to Informational Self-Determination and the Value of Self-Development: Reassessing the Importance of Privacy for Democracy,* in *Reinventing Data Protection?,* Springer, 2009, pp. 61-62.
[56] See CJEU, Joined Cases C-293/12 and C-594/12, Digital Rights Ireland, para 37. See also Tene & Polonetsky (2013), *Big Data for All: Privacy and User Control in the Age of Analytics*, supra note 3, p. 256.
[57] See Tene & Polonetsky (2013), *Big Data for All: Privacy and User Control in the Age of Analytics*, supra note 3, p. 251.
[58] See CJEU, Joined Cases C-293/12 and C-594/12, Digital Rights Ireland, para 27.
[59] Consider the often-cited example concerning the retail chain Target Inc. which could, by analyzing purchasing habits, accurately predict customers' pregnancy and due date. In one case a teenage girl received coupons and advertisements for baby products to her family home, which revealed her pregnancy for her parents. The girl probably not wished for Target Inc. to reveal this information and thus most likely perceived this as a rather serious violation of her privacy. See Charles Duhigg, *How Companies Learn Your Secrets*, The New York Times Magazine, 16 February 2012, http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all, accessed 1 March 2017.
[60] See European Commission, *Special Eurobarometer 431 Data Protection Summary* (2015), <http://ec.europa.eu/public_opinion/archives/ebs/ebs_431_sum_en.pdf>, accessed 28 February 2017.
[61] Critics mean that big data analytics do not only imply mere privacy threats, but also other concerns such as racial or other profiling, discrimination, exclusion, over-criminalization and other restricted freedoms. See Tene & Polonetsky (2013), *Privacy and Big Data: Making Ends Meet,* supra note 10, p. 25. See also Tene & Polonetsky (2013), *Big Data for All: Privacy and User Control in the Age of Analytics*, supra note 3, p. 251.

family life,[62] as well as by the CFREU and the TFEU to ensure its people protection of their personal data.[63] Further, all member states of the EU agreed on 24 October 1995 to adopt a directive on the protection of personal data and the free movement of such data.[64] However, the directive has not succeeded establishing an equal level of protection across the union,[65] and therefore a regulation was adopted on 27 April 2016, which shall apply from 25 May 2018.[66] Both the DPD and the GDPR establish limitations on the processing of personal data and other requirements on actors handling such data.[67] The most central and relevant provisions in the context of big data analytics will be discussed below. It should be noted that these rules are virtually the same in the DPD and the GDPR, although some amendments have been conducted in the new regulation. The following presentation is based on the wording of the provisions in the GDPR, since it is the legal framework that we have to adhere to in the future. However, the corresponding provisions in the DPD are also referred to, since the statute will remain in force for another year.

First of all, certain basic principles are established for the processing[68] of personal data, which an actor, who is considered a controller[69] in relation to the data, has to comply with.[70] The first principle implies that personal data shall be processed in a lawful, fair and transparent way.[71] Secondly, personal data must be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes" ('purpose limitation').[72] Moreover, the personal data being processed must be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are pro-

[62] In the Joined Cases C-465/00, C-138/01 and C-139/01, Rechnungshof v Österreichischer Randfunk and Others, para 21, the CJEU held that the provisions of the DPD must be interpreted in light of the right to privacy stated in article 8 in the ECHR. This statement will remain relevant even after the GDPR has entered into force.
[63] See article 8 in the ECHR, article 7 and 8 in the CFREU, and article 16(1) in the TFEU.
[64] 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.
[65] Recital 9 in the GDPR.
[66] Regulation 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, see article 99.
[67] To clarify, the DPD and the GDPR solely applies to data that is considered personal. See article 3 in the DPD and article 2 in the GDPR. For the definition of 'personal data', see article 2(a) in the DPD and article 4(1) in the GDPR, as well as Section 3.1 below, where this definition is further discussed.
[68] By 'processing' means any operation that is performed on personal data. For the full definition see article 4(2) in the GDPR and article 2(b) in the DPD. See also CJEU, Case C-101/01, Lindqvist, para 25; and Case C-131/12, Google Spain, para 26-31, where it is being discussed what is regarded as processing.
[69] A controller is a "natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data". See article 4(7) in the GDPR and article 2(d) in the DPD. See also CJEU, Case C-131/12, Google Spain, para 33-41, where the definition of controller is being discussed.
[70] See article 5 in the GDPR and article 6 in the DPD.
[71] Article 5(1)(a) in the GDPR and article 6(1)(a) in the DPD.
[72] Article 5(1)(b) in the GDPR and article 6(1)(b) in the DPD.

cessed" ('data minimization').[73] Personal data must also be accurate and kept up to date.[74] Furthermore, personal data shall be "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" ('storage limitation').[75] Lastly, the controller must, according to the new regulation, ensure appropriate security of the personal data being processed, which includes protecting the data against unauthorized and unlawful processing, accidental loss, destruction and damage.[76]

Besides fulfilling all the principles mentioned above, the controller must be able to demonstrate that the processing of the personal data is based on one of the legal grounds stated in the law. In order to be lawful, the data subject concerned must have given his or her consent to the processing, or the processing has to be necessary for some other legitimate purpose.[77] In addition, the controller shall, when collecting personal data from either the data subject itself or another source, provide the concerned data subject with various types of information.[78] Such as, for instance, information regarding the purpose and the legal basis for the processing, potential recipients of the data and the period for which the data will be stored. Another central requirement is that the controller must implement appropriate technical and organizational measures in order to ensure that processing of personal data is performed in accordance with the legislation.[79] To conclude, the principles presented above together with the provisions mentioned in this paragraph require quite a lot from controllers. However, only a few of the requirements, entrenched in the law, have been discussed herein. This indicates that EU data protection legislation places very high demands on actors wishing to process personal data, for the purpose of ensuring a strong protection for people's privacy. Although strong data protection is desirable, one should bear in mind that it may, for reasons explained below, hinder important uses of data.

When it comes to big data analytics, information is rarely gathered for a predetermined purpose, or the information is collected for a specified and explicit purpose but it is later discovered that the information can be used for other purposes.[80] For instance, the data used in

[73] Article 5(1)(c) in the GDPR and article 6(1)(c) in the DPD.
[74] Article 5(1)(d) in the GDPR and article 6(1)(d) in the DPD.
[75] Article 5(1)(e) in the GDPR and article 6(1)(e) in the DPD.
[76] Article 5(1)(f) in the GDPR.
[77] See article 6(1) in the GDPR and article 7 in the DPD.
[78] Article 13 and 14 in the GDPR and article 10 and 11 in the DPD. See generally CJEU, Case C-201/14, Bara and Others, regarding data subjects' right to information.
[79] Article 24 in the GDPR and article 17 in the DPD.
[80] James R. Kalyvas & Michael R. Overly, *Big Data: A Business and Legal Guide*, Auerbach Publications, 2014, p. 33.

Google Flu Trends were from the beginning collected for another purpose than predicting the spread of the H1N1 virus.[81] Hence, it is clear that collected data may be extremely valuable for other purposes than the initial one. However, the propensity of big data to reuse data for different purposes does not fit very well with the purpose limitation principle entrenched in the law.[82] If a controller starts processing data for another purpose that is incompatible with the initial one, the controller is in violation of the law and from 25 May 2018 risk to receive an administrative fine of up to 20 000 000 EUR or 4 % of the total worldwide annual turnover of the preceding financial year.[83] Although further processing for archiving purposes in the public interest, scientific and historical research purposes and statistical purposes should not be considered as incompatible with the initial purposes,[84] most uses of big data do not fall under these exceptions. More commonly data is reused in big data analytics for other purposes than those just mentioned and if further processing is to be lawful in such cases the controller has to obtain a new consent from every individual whose personal data is about to be processed or the processing has to be necessary for some other legitimate reason.[85] In addition, the controller must inform all data subjects concerned about the changed purpose and provide them with any other relevant information.[86] To fulfill these requirements may be very problematic in the context of big data, due to the fact that such enormous datasets often contain personal data relating to an unmanageable amount of people. Hence, the purpose limitation provision certainly makes it difficult to use personal data in big data analytics and thus may prevent new valuable discoveries.

Moreover, the principle of data minimization is also difficult to comply with when conducting big data analytics, since often more data than is necessary for the initial purpose has to be collected in order to find new groundbreaking correlations.[87] Commentators such as Ira Rubinstein have even argued that data minimization requirements are inimical to the underlying thrust of big data and are diminishing the economic as well as social benefits associated with the analysis of such data.[88] Accordingly, also the data minimization provision constrains society from taking advantage of the possibilities with big data.

[81] Tene & Polonetsky (2012), supra note 52.
[82] ICO (2014), supra note 18, p. 40.
[83] Articles 5(1)(b) and 83(5)(a) in the GDPR.
[84] Articles 5(1)(b) and 89(1) in the GDPR. See also article 6(1)(b) in the DPD.
[85] Article 6 in the GDPR.
[86] Articles 13(3) and 14(4) in the GDPR.
[87] See Kalyvas & Overly, supra note 80.
[88] Ira S. Rubinstein, *Big Data: The End of Privacy or a New Beginning?,* International Data Privacy Law, Vol. 3, No. 2, pp. 74-87, 2013, p. 78.

Furthermore, as evidenced by for example the Google Flu Trends case, data can be found to be useful for purposes no one could ever have thought of by the time the data was collected. Hence, in order to discover such new, useful purposes, the data must often be stored for a longer period of time than is considered necessary for the initial purpose. Personal data may, according to the law, be stored for longer periods insofar as the data is only processed for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes.[89] However, as also mentioned in relation to the purpose limitation principle, most processing of personal data within big data analytics takes place for other purposes than those exempted in the law. Accordingly, in most cases longer storage periods than necessary for the initial purpose is not allowed. Hence, the storage limitation requirement does not fit either for big data analytics and thus, it may hamper important uses of data.

It can be concluded that the extensive rules on the protection of personal data entrenched in both the DPD and the GDPR, especially the principles of purpose limitation, data minimization and storage limitation, prevent society from fully harnessing the benefits of big data. As soon as an actor uses personal data for its analysis, the processing of the data falls under the data protection legislation and the actor must comply with the restrictive rules, which virtually make it impossible to conduct big data analytics. Hence, it can be questioned whether the directive, and more importantly the regulation, strikes an adequate balance between beneficial uses of data and privacy risks.[90] At first sight, the legislation seems rather overprotective, leaving no room for utility. However, the enormous amount of personal data being collected, stored, analyzed and shared in the world today may require such strong protection for our privacy. In fact, many people have expressed a fear of what different actors may use their data for and of not having control over their own personal information.[91] Therefore, it would certainly be problematic to reduce the current requirements for processing personal data. To provide a sufficient level of protection for privacy while establishing the right conditions for big data analytics seems nearly impossible. Whether privacy and big data at all can coexist in our increasingly complicated society will be elaborated on in the following chapters.

[89] Articles 5(1)(e) and 89(1) in the GDPR. See also article 6(1)(e) in the DPD.
[90] See generally Eirik Jungar, *Big Data: Mind the Gap – Regulation Meets Reality,* Juridisk Publikation, number 1/2016.
[91] See European Commission, *Special Eurobarometer 431 Data Protection Summary* (2015), supra note 60, accessed 13 December 2016. According to this survey "more than eight out of ten respondents feel that they do not have complete control over their personal data" and "two-thirds of respondents are concerned about not having complete control over the information they provide online".

3. ANONYMIZATION

Although the restrictions upheld by the DPD and the GDPR makes it nearly impossible to reap the benefits of big data, the existence of those restrictions are needed in order to protect everyone's right to privacy. To allow more extensive analysis of big data containing personal information by weakening the protection for our privacy is, as indicated above, not a suitable solution. The question is hence whether it is possible to utilize the benefits of big data without reducing the level of protection for our personal information.

As a matter of fact, big data analytics can unlock mysteries of manufacturing, healthcare, financial markets, cyber security and many more areas without delving into data at an individual level.[92] Many big data actors are not interested in the individuals behind the information, but merely in the information on a more abstract level. If the aim is not to issue targeted advertisements, but rather, for example, to improve a product or service, the actor does not need to know whom certain data belongs to in order to discover correlation and patterns indicating how the product or service in question could be optimized.

Therefore, different anonymization techniques have been developed, which aim is to deidentify the data so as it is no longer considered personal.[93] If the aim is fulfilled the data can be processed freely since it then falls outside the scope of the data protection legislation. Accordingly, at least in theory, anonymization can enable beneficial uses of big data without needing to limit the protection for our privacy. What anonymization is, what different techniques that are available and whether anonymization really is a functioning method for unlocking the benefits of big data will be analyzed below.

3.1 What is Anonymization?

The application of the data protection legislation depends first of all on whether the data in question is considered to be personal or not. Personal data falls under the scope of the legislation, whereas all other data falls outside the scope and thus can be processed freely.[94] The goal of anonymization is to make personal data non-personal, in order to avoid application of the legislation. Hence, it is crucial to first determine what personal data is before moving on to the definition of anonymization.[95]

[92] Ohm (2013), supra note 41, p. 344.
[93] See Hrushikesha Mohanty, Prachet Bhuyan & Deepak Chenthati, *Big Data: A Primer,* Studies in Big Data, Vol. 11, Springer India, 2015, p. 124.
[94] See article 3(1) in the DPD and article 2(1) in the GDPR.
[95] See ICO, *Anonymisation: Managing Data Protection Risk Code of Practice,* November 2012, p. 11, available at www.ico.org.uk.

Personal data is "any information relating to an identified or identifiable natural person" and an identifiable natural person is in this context "one who can be identified, directly or indirectly[96], 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".[97] It can be concluded that the definition of personal data is very broad,[98] which implies that comprehensive measures must be taken in order to anonymize such data, since the broader the definition of personal data is, the harder it is to render such data non-personal.

The distinction between personal and non-personal data as well as the concept of anonymization is being addressed in Recital 26 in both the DPD and the GDPR. There, it is being clarified that the data protection legislation should not apply to anonymous information, meaning information that cannot be related to an identified or identifiable natural person. Regarding the term 'identifiable natural person' it is stated "to determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used … either by the controller or by another person to identify the natural person directly or indirectly".[99] Accordingly, personal data is rendered anonymous according to the law when an individual cannot be identified via all means that is reasonably likely to be used by not only the controller but also any other person. Hence, it can be concluded that both the DPD and the GDPR sets a very high standard on anonymization techniques and places a large burden on those wishing to use anonymized data.[100]

The wording 'all the means reasonably likely to be used' is to be seen as a criterion that shall be applied in order to determine whether an anonymization process is sufficiently robust.[101] Unfortunately this criterion is very vague, which makes it difficult to ascertain whether an anonymization technique lives up to the requisite standard. Luckily, the newly adopted GDPR, unlike the DPD, gives further guidance on what 'all the means reasonably likely to be used' entails. It is stated that, when determining whether means are reasonably likely to be used in order to identify an individual, "account should be taken of all objective factors, such

[96] The word 'indirectly' should be interpreted as meaning that it is not necessary that the information under consideration alone allows the data subject to be identified, in order to treat the information as personal data. See CJEU, Case C-582/14, Breyer, para 41.
[97] Article 4(1) in the GDPR. See article 2(a) in the DPD for a slightly different definition. See also CJEU, Case C-582/14, Breyer, para 49; Case C-291/12, Schwarz, para 27; and Case C-342/12, Worten, para 19, regarding what could constitute personal data.
[98] A29WP, Opinion 4/2007 on the Concept of Personal Data, p. 4.
[99] The quotation is from Recital 26 in the GDPR. Recital 26 in the DPD varies with regards to a few words, but the essential meaning of the wording is the same.
[100] A29WP, Opinion 05/2014 on Anonymisation Techniques, p. 6.
[101] Ibid., p. 8.

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".[102] Even though this indicates which factors to consider when determining whether certain data is personal or not, it can be concluded that neither the current nor the forthcoming data protection legislation within the EU provides us with a clear definition of anonymization.

Moreover, what is further interesting is that account should be taken of the means reasonably likely to be used by either the controller or by 'another person'.[103] What this really entails is subject to discussion. In fact, there are two possible ways of interpreting Recital 26, by academics referred to as the "absolute/objective approach" and the "relative/subjective approach".[104] According to the former approach data is considered to be personal if any third party is able to determine the identity of the individual, while the latter approach would treat certain data as personal data only if the controller has the legal and practical means of obtaining the additional information necessary for identifying the individual from a third party.[105] The wording in Recital 26 suggests the absolute approach.[106] This approach has also been supported by the A29WP.[107] However, in a rather recent decision by the CJEU, commonly referred to as *Breyer*, the Court embraced the relative approach.[108] The Court first held that "for information to be treated as 'personal data' … it is not required that all the information enabling the identification of the data subject must be in the hands of one person".[109] However, the Court further stated that for data to be considered as personal, the possibility to combine that data with the additional identifying information held by a third party must constitute a means likely reasonably to be used, and "that would not be the case if the identification of the data subject was prohibited by law or practically impossible on account of the fact that it requires a disproportionate effort in terms of time, cost and man-power, so that the risk of identification appears in reality to be insignificant.[110] Based on these statements it is clear that the Court adopted the relative approach in this case. However, as Ira Rubinstein has conclud-

[102] Recital 26 in the GDPR.
[103] See Recital 26 in the GDPR. In Recital 26 in the DPD the term 'any other person' is being used. Assumingly, the two terms have the same meaning.
[104] Frederik J. Zuiderveen Borgesius, *Singling Out People Without Knowing Their Names – Behavioural Targeting, Pseudonymous Data, and the New Data Protection Regulation*, 32 Computer Law & Security Review 256, 2016, pp. 263-265. See also Ira Rubinstein, *Framing the Discussion*, Brussels Privacy Symposium on Identifiability: Policy and Practical Solutions for Anonymisation and Pseudonymisation, 2016, p. 8, available at https://fpf.org/brussels-privacy-symposium-final-papers/.
[105] Rubinstein (2016), supra note 104.
[106] Ibid.
[107] A29WP (2014), supra note 100, p. 9.
[108] CJEU, Case C-582/14, Breyer.
[109] Ibid., para 43.
[110] Ibid., para 45-46.

ed, the Court did not explicitly reject the absolute approach.[111] Therefore, we will have to await further statements by the CJEU before we can know for sure if the relative approach should be applied in all cases or merely under certain circumstances.

Having to adhere to the absolute approach instead of the relative approach can lead to severe consequences for big data actors. This can be illustrated by an example taken by the A29WP.[112] Suppose that a dataset containing personal data undergoes an anonymization process. Further suppose that the controller conducting the anonymization does not delete the original identifiable data. In this case, the "anonymized" data is still personal data in relation to not only the controller, but also to any other party who receives the dataset, regardless of whether such a party has the means to access the original data and actually identify the data subjects. Whereas according to the relative approach, the data is only personal in the hands of an actor who has the legal and practical means to get access to the raw data. In a broader perspective, this means that if the absolute approach is being applied, an actor must control that no one worldwide has the means to identify any data subject within a dataset before that actor can treat the dataset as anonymous. In contrast, if the relative approach is being applied, such as in the *Breyer* case, an actor will only have to determine whether the actor itself has the means to identify an individual within the dataset in order to know if the dataset falls within or outside the scope of the legislation. Clearly there is a significant difference between the absolute and the relative approach, and in order to get a clear picture of what really constitutes personal data according to the legislation and thus what is required with regards to anonymization, it is crucial that the CJEU gives further guidance on how the interpret Recital 26.

Both the fact that the wording 'all the means reasonably likely to be used' is very vague, and the fact that it is rather unclear whether the relative approach should be applied in all cases or the absolute approach should be adopted in some, makes it difficult to determine when data is regarded as anonymous according to the law. It can hence be concluded that neither the DPD nor the GDPR has any clear legal boundaries.[113] This implies that it certainly is problematic to determine beforehand whether a specific anonymization process fulfills the criterion of anonymization entrenched in the law, since this criterion is somewhat unclear. Accordingly, an actor can never be certain that the data will be regarded as anonymous in case of a review and thus can never know for sure whether the data will fall within or outside the scope of the legislation.

[111] Rubinstein (2016), supra note 104.
[112] See A29WP (2014), supra note 100, p. 9.
[113] This is being further discussed in Section 3.3 and Chapter 4 below.

In recent years it seems to have become even more problematic to use anonymization as a method to reap the benefits of big data. Computer scientists have proven, in several cases, that anonymized data can easily be reidentified.[114] It has even been argued that a risk of reidentification is inherent to anonymization.[115] This revelation raises the question whether we at all should continue to put our trust in anonymization for the purpose of enabling broader processing. In order to answer that question, one must evaluate whether existing anonymization techniques, despite the increased possibilities to reidentify data, are sufficient for exempting data from the scope of the legislation. Due to the fact that these anonymization techniques are very different, it is necessary to examine each of them in great detail, to be able to determine whether a certain technique is sufficient or not.

However, before this examination can be conducted it must be clarified what an anonymization technique has to live up to in order to exempt data from the scope of the legislation. As concluded above, data is considered anonymous according to both the DPD and the GDPR, and thus falls outside the scope of these statutes, when the individual behind the data cannot be identified by using all means reasonably likely to be used by either the controller or another person to identify that individual. In other words, it should be reasonably impossible to identify the individuals in a dataset, in order for that dataset to be considered anonymous.[116] It should be noted that this criterion remains the same regardless of whether Recital 26 should be interpreted according to the relative or the absolute approach. Hence, the anonymization techniques presented below will be evaluated with regards to whether they live up to the criterion of making identification reasonably impossible.

3.2 Different Anonymization Techniques

Neither the DPD nor the GDPR provides any guidance on how anonymization should be carried out.[117] Therefore, different anonymization techniques have been developed within the technological field, which are more or less robust against re-identification. A few different techniques, which are commonly used, will be presented and discussed herein.

---

[114] See for example Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization,* 57 UCLA Law Review 1701, 2010, p. 1701. See also Section 3.2 and 3.3 below.
[115] A29WP (2014), supra note 100, p. 7.
[116] The term 'reasonably impossible' has also been used by the A29WP to describe the standard an anonymization process has to live up to in order to be considered sufficiently robust. See A29WP (2014), supra note 100, p. 8.
[117] In Recital 26 in the DPD it is stated "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". However, no further guidance is given in the legislation as to which techniques to apply. Recital 26 in the GDPR does not even encourage the adoption of codes of conduct. See also ICO (2012), supra note 95, p. 12.

Before delving into the characteristics of these techniques, one must first understand the difference between so-called direct identifiers and quasi-identifiers. Direct identifiers are information that directly identifies a natural person, such as name and social security number.[118] Hence, the first step when trying to deidentify individuals in a dataset is to remove these direct identifiers, which can be done either by replacing them with random values or with pseudonyms.[119] However, it is often not enough to remove direct identifiers in order to make it reasonably impossible to reidentify data subjects.[120] For instance, if a dataset contains information such as home address and age, a third party can simply by using a search engine on the web see who is living on the address and by the given age often identify the individual in the dataset.[121] Hence, it is clear that identification is often possible even though the dataset only contains quasi-identifiers, such as for example address and age. Quasi-identifiers cannot, unlike direct identifiers, in itself identify a natural person. However, they can be linked together with other information, as seen in the example taken above, so as an identification of an individual is possible.[122] Therefore, in most cases also quasi-identifiers need to be removed in order to obtain an anonymized dataset. However, this is more challenging than it seems. Quasi-identifiers often constitute important information and can be crucial for the analysis of the data.[123] Hence, it is often problematic to simply delete such information.

Therefore, several techniques have been developed that allows big data actors to preserve important quasi-identifiers in the dataset, while making it significantly more difficult to identify data subjects based on these identifiers. Such techniques, referred to as anonymization techniques, thus primarily target quasi-identifiers and not direct identifiers, since direct identifiers must in most cases be completely removed to accomplish an anonymized dataset.[124] The names of the different techniques and the way the techniques are presented vary in scholarly

---

[118] Ira S. Rubinstein & Woodrow Hartzog, *Anonymization and Risk,* 91 Wash. L. Rev. 703, 2016, p. 710.

[119] Ibid. Pseudonymization, which consists of replacing the most identifying field within a data record with a more artificial identifier (e.g. replacing the name Marco with Loqfh), is not considered to be a method of anonymization and thus will not be discussed further in this thesis. Regarding pseudonymization see A29WP (2014), supra note 100, pp. 20-23.

[120] Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets,* 2008 IEEE Symposium on Security and Privacy, pp. 111-125, IEEE, May 2008, p. 111.

[121] This can be done by using for example the Swedish search engine www.hitta.se or the British search engine www.192.com.

[122] Rubinstein & Hartzog, supra note 118, p. 711.

[123] Ibid., p. 712.

[124] To completely remove an identifier is called suppression, which is also being classified as an anonymization technique. Although suppression provide strong privacy protection and if aggressively applied, in most cases, actually achieve truly anonymized datasets, the technique will not be further presented below for the reason discussed above, that quasi-identifiers constitutes important information and suppressing them often renders the data useless for research. Thus, suppression is not a suitable technique for enabling actors to maximize the opportunities with big data and hence is not of interest in the context of this thesis. Regarding suppression see Ohm (2010), supra note 114, pp. 1713-1714. See also Rubinstein & Hartzog, supra note 118, p. 712.

works. However, the characteristics of the different techniques are the same. The below presentation follows to a great extent the structure in the A29WP's opinion on anonymization techniques, which divide such techniques into two different families, the first one being data randomization and the second one being data generalization.[125] Under each of these families several techniques are subordinated. There are generally speaking three established randomization techniques, which are noise addition, permutation and differential privacy, as well as three generalization techniques, referred to as k-anonymity, l-diversity and t-closeness. These six anonymization techniques will be presented and analyzed below.

In order to understand the presentation of the different techniques, a few terms need to be clarified. All anonymization techniques presented below are designed to deidentify personal data in structured materials. Typically, personal data that is going to be used for research are stored in tables. Hence, each technique is presented as if it was applied to a table with personal data (see for example the illustration below), where the whole table is often referred to as the dataset. Each row in a table corresponds to one single data subject[126] and is called a record. Each record or row is composed of a set of values (e.g. 1956, male, female etc.) for a number of attributes (e.g. year of birth, gender, zip code, yearly income, disease etc.). All these attributes can be quasi-identifiers. Attributes such as income and disease are referred to as sensitive attributes, since if revealed may cause significant harm to individuals.[127] Furthermore, in the following, two types of information disclosure will be discussed, identity disclosure and attribute disclosure. The former refers to when a specific individual can be identified, and the latter refers to when one can link a value of an attribute to an individual, but where identification is not possible. Any other terms used below, which need explanation, will be clarified in a footnote or in the mere text.

Year of birth	Gender	Zip code	Yearly income (EUR)
1956	M	113 54	80 000
1956	F	114 56	200 000
1994	F	104 05	20 000

[125] A29WP (2014), supra note 100, p. 10.
[126] The terms data subject and individual are used interchangeably.
[127] It should be noted that the word 'sensitive' is used within this thesis to refer to such information that generally is perceived as sensitive and thus may cause harm to individuals if revealed. Accordingly, the word is given a broader meaning within this work than it is given in the legislation, where the term 'sensitive data' merely refers to a number of special categories of personal data. See Recital 10 and 51, and article 9 in the GDPR, as well as article 8 in the DPD.

3.2.1 Randomization

Before considering the specific characteristics of each technique, some general comments on randomization are necessary. By simply reading the word randomization one can get a glimpse of what such techniques entails. The goal of these techniques is to de-identify personal data by randomize the data. In other words, what randomization techniques have in common is that they alter the veracity of the data in order to remove the connection between the data and the individual.[128] However, the technical measures that are applied to the data vary between the different randomization techniques, which will be demonstrated below.

3.2.1.1 Noise Addition

Noise addition is a randomization technique that modifies certain attributes within a dataset so as they become less accurate.[129] In practice, this means that in a dataset containing personal data about people's weight, the entries will only be accurate to for example +-10 kilos. Although the accuracy of the data is being decreased, the intent is to retain the overall distribution in the dataset. How much the values of the attributes need to be modified will depend on how sensitive the data is verses how accurate the information have to be in order to be useful.[130] The aim when applying the technique is of course to unable direct identification of the individuals in the dataset, but also to make it impossible to determine how the data have been modified and thus to repair the data.[131] However, if the values of the attributes in the dataset vary significantly it may be possible to determine which data that belongs to who through cross-correlation with information in other databases or with the help of background knowledge.[132] It may also be possible to repair the data by logic reasoning if the noise added is inconsistent or out-of-scale.[133] In these cases, it is certainly not reasonably impossible to identify an individual within a dataset. Thus, noise addition is not always a feasible technique that invariably fulfills the goals of anonymizing personal data and enabling broader processing. This can be illustrated by the following commonly cited example.

In 2006, Netflix, which is one of the world's largest video content providers, announced a contest, called the Netflix Prize, that sought to improve their movie recommendation ser-

[128] A29WP (2014), supra note 100, p. 12.
[129] Ibid.
[130] Ibid.
[131] Ibid.
[132] If the records in a dataset are based on very rare individual attributes, that dataset is said to be too sparse. This sparsity is empirically well established and especially common in datasets covering information regarding individual transactions and preference records. See Narayanan & Shmatikov (2008), supra note 120.
[133] A29WP (2014), supra note 100, p. 13.

vice.[134] To enable the participants to develop suitable algorithms for this service Netflix publicly released a dataset containing over 100 million ratings on over 18 000 movies, expressed by almost 500 000 Netflix users during the period between October 1998 and December 2005.[135] The dataset had been "anonymized" according to Netflix's internal privacy policy and was said to have been stripped of all personal information identifying individual customers.[136] Although noise was added, as ratings and dates of ratings had been altered, it turned out that users could easily be identified with the help of rather limited background knowledge.[137] Arvind Narayanan and Vitaly Shmatikov proved in their article *Robust De-anonymization of Large Sparse Datasets* that users could be identified by linking the records in the dataset with information publicly available on the Internet Movie Database (IMDb).[138] It is demonstrated that an actor only needs to have a little background knowledge in order to single out a user or at least identify a small set of records that include a specific user's ratings.[139] It is further shown that the background information does not even have to be precise. Rating dates do only need to be known with a 14-day error and ratings may only be known approximately.[140] In some cases identification was shown to be possible even if the ratings and dates had been altered so as they were completely wrong.[141] By identifying which ratings that had been expressed by who, Narayanan and Shmatikov obtained knowledge about users' political preferences and other potentially sensitive information.[142] Hence, this example shows that noise addition may fail to live up to the requisite standard of anonymization stated in the DPD and the GDPR and thus may not be relied on for the sake of unlocking the benefits of big data.

3.2.1.2 Permutation

Permutation, or swapping, is as noise addition also a randomization technique.[143] However, instead of modifying the data, permutation preserves each value of the attributes but shuffles the values between different records, which lead to that some of the values are connected to the "wrong" data subject.[144] The idea is that it should be impossible to determine which val-

[134] See the official website for the contest, www.netflixprize.com, accessed 31 January 2017.
[135] Ibid., under the heading 'Rules'.
[136] Ibid. See also A29WP (2014), supra note 100, p. 13.
[137] Narayanan & Shmatikov (2008), supra note 120, pp. 111 and 119.
[138] Ibid., pp. 112-113.
[139] Ibid., p. 112.
[140] Ibid.
[141] Ibid.
[142] Ibid., pp. 111 and 123.
[143] Rubinstein & Hartzog, supra note 118, p. 712. A29WP (2014), supra note 100, p. 13.
[144] A29WP (2014), supra note 100, p. 13.

ues that belong to which individual. This technique is suitable when it is crucial for the analysis that each value of the different attributes remains exact, but where it is less important to be able to link a value with the characteristics of a certain individual.[145]

An example may be useful. Lets say that a dataset contains personal data of people living in a certain city in terms of age, zip code and yearly income. If permutation is applied to this data, some of the values of the different attributes will be swapped, so that a record that before the technique was applied showed that an individual at the age of 20, who live in an area with the zip code 104 05, earn 20 000 euros per year, will after the swapping show that the same individual earn for instance 200 000 euros per year. Whereas another record that originally showed that a 60 year old, with the zip code 114 56, earn 200 000 euros, will after the permutation technique has been applied show that the 60 year old has an annual earning value of 20 000 euros. As illustrated, the values of the attributes remains the same but are swapped between different records, in order to prevent identification of the individuals in the dataset. If the data shall be used for calculating the average yearly income in the whole city, this anonymization technique can be used without affecting the analysis. However, if the data is to be used for determining the average yearly income within different residential areas and its connection with the average age in that area, then permutation will diminish the value of the data. Hence, permutation is only useful for certain types of data analysis and in cases as the one described above, attributes such as age and zip code may just as well be deleted if the goal is to merely determine the annual income of the entire city.[146]

Moreover, permutation tends not to be particularly robust against re-identification attempts. When it comes to personal data, attributes often have strong logical relationships, such as in the example taken above where age, income and residential area often are correlated. In datasets with such closely correlated attributes, permutation will most likely not achieve an anonymized dataset according to the law, since it is relatively simple to reverse the permutation through inference.[147] Although one cannot determine with absolute certainty which values of the different attributes that belong to a specific individual, a probabilistic inference is often enough to enable a reidentification of an individual by linking the values to information in other databases.[148] Hence, it can be concluded that neither permutation is an anonymization technique that can be fully relied on by big data actors, at least not in the case where the attributes in the dataset are strongly correlated.

[145] Ibid.
[146] Regarding suppression see supra note 124.
[147] A29WP (2014), supra note 100, p. 14.
[148] Ibid.

3.2.1.3 Differential Privacy

Differential privacy is a concept originated by the American computer scientist Cynthia Dwork.[149] The concept differs significantly from the two previously presented techniques, but can still be subordinated under the family of randomization techniques, since it implies adding noise to the data. As will be evidenced below, several anonymization techniques are robust against various types of reidentification attacks, but can, however, not guarantee that personal information can be inferred about a data subject if someone possesses all background knowledge needed in order to link a certain value of an attribute to the right individual.[150] Suppose that a dataset contains information about the average height of Lithuanian women. Further suppose that an actor who is aiming to reidentify an individual, referred to as an adversary within computer science, knows that this individual is two centimeters shorter than the average Lithuanian woman. The so-called adversary can by combining this background knowledge with the information in the dataset, learn the height of the targeted individual.[151] Hence, even if the dataset has been anonymized to such an extent that information cannot be linked to a specific individual by just having access to that dataset, an adversary with the right background knowledge can still infer information about a certain individual. Cynthia Dwork argues that it is impossible to develop an anonymization technique that can prevent an adversary with arbitrary amounts of background knowledge to retain information about an individual from a released dataset.[152] This is because an adversary can infer such information, as the height of the targeted individual described above, regardless of whether that individual is in the dataset or not. Hence, the goal of differential privacy is therefore not to prevent such inference attacks (because this is impossible), but to ensure that the risk to one's privacy is not substantially increased as a result of that one's record is contained in a released dataset.[153] Accordingly, any negative effect with regards to one's privacy, caused by the release of the dataset, are not due to the individual's presence in the dataset but due to other factors, such as the adversary's prior knowledge.[154] In other words, the harm caused to an individual by the

[149] See Cynthia Dwork, *Differential Privacy,* in *Automata, Languages and Programming*, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II, Springer Berlin Heidelberg, 2006.
[150] See Sections 3.2.2.1 and 3.2.2.2 where this problem is discussed.
[151] This is the example taken by Cynthia Dwork to illustrate the problem. See Dwork (2006), supra note 149, p. 2.
[152] Ibid.
[153] Ibid.
[154] Felix T. Wu, *Defining Privacy and Utility in Data Sets,* 84 University of Colorado Law Review 1117, 2013, p. 1138.

release of the dataset will be essentially the same independent of whether that individual's data is in the dataset.[155]

The question is then how differential privacy can be achieved. Cynthia Dwork suggests that it should be achieved in an interactive rather than in a non-interactive setting.[156] Thus, Dwork presents an interactive privacy mechanism in which the data controller answers questions about the data, instead of publishing a deidentified version of the entire dataset.[157] To achieve differential privacy the data controller should add appropriately chosen random noise to the answer.[158] Hence, any data released have been altered so that no answers provided are completely accurate. How much noise that needs to be added depends on how much the answer to a question changes when an individual's data changes.[159] This means that a query about a specific individual's value to a certain attribute will receive a very noisy answer, since the original answer would change completely if that individual's data changed.[160] However, a query regarding a larger group of people will receive a less noisy answer. If the presence or absence of an individual's data does not change the true answer at all, then the delivered answer will be very close to the original answer.[161] Hence, in this way differential privacy protects every single individual's right to privacy while still preserving the overall utility of the dataset. The mechanism serves uses of data that concerns a larger group of people, such as a whole population, but prevents attacks on specific individuals and small groups of people. Thus, it can be concluded that differential privacy strikes a rather reasonable balance between the two opposite interests – privacy and utility.

Although differential privacy has several advantages, it also has some limitations. First of all, when differential privacy is achieved in an interactive setting, the data controller has to continuously keep track of the queries received and the answers delivered, in order to avoid that an actor poses multiple questions, which answers can then be combined to disclose in-

[155] Cynthia Dwork, *A Firm Foundation for Private Data Analysis,* 54 Communications of the ACM 86, 2011, p. 91. See also Andrew Chin & Anne Klinefelter, *Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study,* 90 N.C. L. Rev. 1417, 2012, p. 1417.
[156] Dwork (2006), supra note 149, p. 3. Differential privacy can be achieved in a non-interactive setting as well, which after the data has been released does not require any further participation from the data controller. However, it is more difficult to achieve differential privacy in a non-interactive setting and not as many questions can be answered as in an interactive setting. See also Wu, supra note 154, pp. 1138-1139. Due to its limited usefulness, differential privacy as a non-interactive privacy mechanism will not be discussed further herein.
[157] Dwork (2006), supra note 149, pp. 3 and 9. To only release a part of a dataset, which corresponds to a researcher's need, instead of releasing the entire dataset, is also called aggregation. Regarding aggregation see for example Ohm (2010), supra note 114, p. 1715.
[158] Dwork (2006), supra note 149, p. 9.
[159] Ibid., pp. 9-10.
[160] Wu, supra note 154.
[161] Ibid.

formation about an individual.[162] Moreover, if an actor receives answers to enough questions, that actor may even be able to reconstruct the entire dataset, meaning that personal data about every single individual within that dataset can be obtained.[163] Thus, in order to make it reasonably impossible to identify an individual within a dataset with the help of differential privacy, a data controller must not provide answers to too many queries. At some point it will be considered as possible to identify an individual using all means reasonably likely to be used by either the controller or another person. Hence, differential privacy will only fulfill the legal criterion for anonymization insofar as answers to questions are provided which when combined does not allow an actor to identify an individual directly or indirectly.

Moreover, what is particularly important to notice is that differential privacy can only be used by an actor who wants to release data to others, which would otherwise not be considered legally under the DPD and the GDPR. To clarify, the interactive privacy mechanism described above is designed merely for releases of data from a data controller to a third party. Accordingly, differential privacy cannot be used to anonymize personal data to enable big data analytics for internal purposes, such as improving one's product or service. Hence, it can be concluded that differential privacy is only sufficient for certain purposes, but for those purposes it may be a useful technique to employ.[164]

It should further be noted that according to Dwork's interactive mechanism, the controller retains the original dataset with the personally identifiable information and then alters the data before it is released to a third party.[165] This means that the data remains personal in the hands of the controller. Whether the released data can be considered as anonymous in the hands of a third party depends on if the absolute or the relative approach is adopted.[166] According to the absolute approach, Dwork's differential privacy mechanism does not render the data anonymous in relation to a third party insofar as the controller retains a copy of the original data.[167] Hence, if Recital 26 is interpreted according to the absolute approach, the controller will have to erase the original data, if the released data is to be rendered anonymous with the help of differential privacy. However, if the relative approach is applied, the released data can be re-

[162] A29WP (2014), supra note 100, p. 16.
[163] Wu, supra note 154, p. 1140.
[164] In an overview of the problems of reidentification and potential solutions, Arvind Narayanan and Vitaly Shmatikov endorse differential privacy as a "major step in the right direction," but concede that it is not adaptable to all situations and must be "built and reasoned about on a case-by-case basis." Arvind Narayanan & Vitaly Shmatikov, *Privacy and Security: Myths and Fallacies of 'Personally Identifiable Information'*, 53 Communications of the ACM 24, 2010, pp. 24-26.
[165] A29WP (2014), supra note 100, p. 15.
[166] Regarding the absolute and the relative approach see Section 3.1 above.
[167] See A29WP (2014), supra note 100, p. 9.

garded as anonymous in relation to a third party even if the controller retains the original data, provided that the possibility to combine the released data with the original data held by the controller does not constitute a means reasonably likely to be used.[168] Since the CJEU embraced the relative approach in the recent *Breyer* case, differential privacy may become more important for big data analytics in the future.

To summarize, differential privacy may render data anonymous according to the legislation, provided that the answers released to a third party does not enable that party to identify a data subject directly or indirectly. Hence, a data controller must keep track of the queries received and the answers provided in order to make identification reasonably impossible. Furthermore, as concluded above Dwork's differential privacy mechanism can only render personal data anonymous if Recital 26 is interpreted according to the relative approach. If the absolute approach is adopted, the original data has to be erased for differential privacy to be of any value for big data actors. We will have to await further guidance from the CJEU before we can determine how differential privacy should be conducted in order to exempt data from the scope of the legislation.

3.2.2 Generalization

Besides randomization techniques have so called generalization techniques been developed. What all generalization techniques have in common is that they generalize, or dilute, the attributes or the values in a dataset.[169] The result is that more general categories replace more specific information about data subjects, which makes the individuals harder to identify.[170] The generalization techniques that will be presented below are all based on the same idea but vary in such a way that they are refinements of each other.

3.2.2.1 k-Anonymity

k-Anonymity is a concept originally established by Latanya Sweeney and Pierangela Samarati.[171] The concept is primarily applied on datasets in order to prevent an adversary from being able to single out an individual's record, i.e. being able to link the values of each attribute to

[168] See CJEU, Case C-582/14, Breyer, para 45.
[169] A29WP (2014), supra note 100, p. 16.
[170] Wu, supra note 154, p. 1133.
[171] See Latanya Sweeney & Pierangela Samarati, *Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression,* IEEE Security and Privacy, 1998; Latanya Sweeney, *k-Anonymity: A Model for Protecting Privacy,* International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002, pp. 557-570; Latanya Sweeney, *Achieving k-Anonymity Privacy Protection Using Generalization and Suppression,* International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002, pp. 571-588.

the correct data subject.[172] The technique aims to prevent this by grouping each data subject with at least *k* other individuals.[173] In other words, *k*-anonymity implies that, given what the adversary already knows, the adversary can never narrow down the set of potential target records to fewer than *k* records in the dataset.[174] Hence, the data have to be altered in such a way that a value of an attribute, for instance 1956 as a year of birth, appears at least *k* times in the dataset.[175] What *k* should be determined to will depend on how many times each value of all attributes has to appear in order for it to be reasonably impossible to identify any individual in the dataset.

k-Anonymity is generally achieved by generalizing the values of the attributes in a dataset.[176] The original values are replaced by less specific, more general values.[177] However, the replacing values are still faithful to the original data.[178] For instance, attributes like date of birth and zip code can be changed to year of birth and city, which will automatically lead to that a more general value will replace a more specific and thus a larger group of data subjects will have the same value in response to these attributes. Further, if the attributes are not suited for being generalized, the values may be diluted instead. For example, if a dataset contains attributes such as age and height, the exact values of these attributes can be generalized to interval values, e.g. age 20-30 years and height 170-180 cm. In this way sensitive and rare values, which otherwise allows for a rather easy identification, can be protected.[179] These generalization measures can also be combined with suppression, which implies not releasing a value at all.[180] For instance, in a dataset containing the attribute 'race', values such as black, white and Asian can in a few records be suppressed and replaced with 'person'.[181]

It can be concluded that *k*-anonymity provides a solution to the problem, discussed in Section 3.2.1.2, with adversaries being able to single out individuals due to the fact that the dataset contains rare attributes, which can be linked to a data subject by inference or with the

[172] A29WP (2014), supra note 100, pp. 11 and 16.
[173] Ibid., p. 16. A group of *k* individuals is referred to as an equivalence class below, which in other words can be defined as a set of records that have the same value for a certain attribute.
[174] Wu, supra note 154, p. 1142.
[175] Sweeney (2002), *k-Anonymity: A Model for Protecting Privacy*, supra note 171, pp. 564-565.
[176] A29WP (2014), supra note 100, p. 16.
[177] Sweeney (2002), *Achieving k-Anonymity Privacy Protection Using Generalization and Suppression,* supra note 171, p. 575.
[178] Ibid.
[179] Narayanan & Shmatikov (2008), supra note 120, p. 112.
[180] Sweeney (2002), *Achieving k-Anonymity Privacy Protection Using Generalization and Suppression,* supra note 171, p. 572. Regarding suppression see also supra note 124.
[181] Sweeney (2002), *Achieving k-Anonymity Privacy Protection Using Generalization and Suppression,* supra note 171, p. 582.

help of background knowledge.[182] Furthermore, *k*-anonymity is a suitable technique for big data actors in such a way that the values are just being generalized, and not modified so as they become inaccurate which is the case with noise addition, and hence are still faithful to the original data. However, even if *k*-anonymity offers advantages, several scholars have accused the technique for being insufficient.[183]

Although *k*-anonymity prevents an adversary from singling out a targeted individual, it is still possible to identify an individual by inference.[184] If an adversary knows that a targeted individual is in a dataset and has background information regarding certain quasi-identifiers, the adversary will be able to determine which group of *k* individuals (i.e. equivalence class) the targeted data subject belongs to and thus may retrieve the personal data of interest.[185] This can be illustrated by an example. If a dataset contains the attributes 'year of birth', 'gender', and 'disease', and all individuals that are grouped together by year of birth, e.g. 1956, have the same value with regards to disease, e.g. cancer, then an adversary that knows that the targeted individual is in the dataset and is born 1956 can conclude that the individual has cancer.[186] Hence, it can be concluded that *k*-anonymity does not prevent an adversary from obtaining personal information about an individual if there is no diversity in the values of the attributes within an equivalence class.[187]

Moreover, even if there is some diversity among the values an adversary may still, with the help of background knowledge, be able to conclude with great probability which value that belongs to the targeted individual.[188] For example, suppose that it is determined that the targeted individual belongs to a group of four data subjects, where one data subject has the value 'viral infection' in response to the attribute 'disease' and the rest of the data subjects have the value 'coronary heart disease' in response to the same attribute. If the adversary in such a case knows that the targeted individual is Japanese and it is well known that Japanese have an extremely low incidence of heart diseases, then the adversary can with near certainty conclude

[182] See A29WP (2014), supra note 100, p. 16.
[183] See for example Charu C. Aggarwal, *On k-Anonymity and the Curse of Dimensionality,* Proceedings of the 31st International Conference on Very Large Data Bases (VLDB) 901, 2005, p. 901. See also Wu, supra note 154, pp. 1142-1143; Rubinstein & Hartzog, supra note 118, p. 712; Narayanan & Shmatikov (2008), supra note 120, pp. 112 and 124.
[184] Aggarwal, supra note 183.
[185] A29WP (2014), supra note 100, p. 17.
[186] See Wu, supra note 154, p. 1142. See also Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke & Muthuramakrishnan Venkitasubramaniam, *l-Diversity: Privacy Beyond k-Anonymity*, ACM Transactions on Knowledge Discovery from Data, Vol. 1, No. 1, Article 3, March 2007, pp. 3-4, where this problem is called the homogeneity attack.
[187] Machanavajjhala et al., supra note 186, p. 4.
[188] Ibid., where this problem is called the background knowledge attack.

that the targeted individual has a viral infection, and not a heart disease.[189] Hence, although a group of *k* individuals has some diversity with regards to an attribute, an adversary can still obtain sensitive information by inference based on background knowledge.

The problems with *k*-anonymity, mentioned in the two previous paragraphs, lead to the conclusion that the value of *k* is crucial for whether a dataset will be rendered anonymous by the application of the technique. The smaller the value of *k* is, the more likely is it that individuals can be identified by inference due to the lack of diversity in attributes or based on background knowledge.[190] However, the higher the threshold of *k* is, the more will the values have to be generalized, which will lower the utility of the data.[191] Hence, to determine a value of *k* that makes it reasonably impossible to identify a data subject within a dataset, which not requires the values to be generalized to such an extent that the data becomes useless, seems rather problematic. Accordingly, it can be questioned whether *k*-anonymity is the technique that will enable the society to utilize the benefits of big data while preserving the right to privacy.

3.2.2.2 l-Diversity

As discussed in the previous section, *k*-anonymity does not prevent an adversary from discovering values of attributes when there is no diversity in those attributes, and even if there is little diversity an adversary can often with great probability determine which of the values that belong to the targeted individual based on obtained background knowledge. *l*-Diversity addresses these two problems by requiring that in each group of *k* individuals a sensitive attribute has at least *l* well-represented values.[192] Hence, in an *l*-diverse dataset an adversary can never retain a sensitive value by simply determine which group of individuals the targeted data subject belong to, since all of these individuals will never share the same sensitive value. In other words, if the value of *l* is at least 3 in a dataset containing the attributes 'year of birth' and 'disease', an adversary who knows that the targeted individual is born 1956 and hence is grouped with for instance four other individuals can never determine which disease the targeted individual has, since there are at least three different values (e.g. cancer, coronary heart disease and virus infection) to the attribute 'disease' in this group of four data subjects.

Moreover, *l*-diversity also prevents an adversary from obtaining sensitive information about an individual with the help of background knowledge. For example, even if an adver-

[189] Ibid.
[190] A29WP (2014), supra note 100, p. 17.
[191] Regarding this problem see Aggarwal, supra note 183.
[192] Machanavajjhala et al., supra note 186, p. 16.

sary knows that the targeted individual is born 1956 and hence belongs to a group of four data subjects, and in addition knows that the targeted individual is Japanese and thus is extremely unlikely to have coronary heart disease, the adversary can still not determine whether the targeted individual has cancer or virus infection. However, *l*-diversity can unfortunately not guarantee that an adversary never will be able to retain sensitive information about an individual based on background knowledge, since it is impossible to foresee what a potential adversary may know.[193] In spite of this, an actor wishing to anonymize a dataset can still by employing the *l*-diversity technique determine how difficult it would be for an adversary to infer sensitive information about an individual in the dataset. Such determination can be done since if the actor knows that there are *l* different values to every sensitive attribute in each group of *k* individuals, the actor also knows that an adversary needs $l - 1$ damaging pieces of background information to exclude all possible values except one and thus to be able to identify the value belonging to the targeted individual.[194] Hence, *l*-diversity can guard against many attacks even though it cannot be determined what kind of knowledge an adversary possesses.[195]

Besides the fact that *l*-diversity cannot ensure that an adversary never will be able to withdraw new information about an individual due to extensive background knowledge, the technique has two other shortcomings. First of all, *l*-diversity is unsuitable for datasets that contains one sensitive attribute, which only has two possible values, and where the distribution of these two values is very different.[196] For example, a dataset containing test results for human immunodeficiency virus (HIV), will only have two values in response to that attribute - positive or negative, and will probably have a very low percentage of the former value (e.g. 1%), while the rest of the records (99%) will contain the latter value. In this case, most equivalence classes will contain only the value negative and hence it is inappropriate to try to achieve 2-diversity, since it would lead to a great information loss.[197] Moreover, it would certainly not result in a version that is faithful to the overall distribution in the original dataset. Suppose that when 2-diversity is being achieved one equivalence class ends up containing as many positive values as negative. From this an adversary can draw the conclusion that the individuals in this equivalence class have 50% possibility of being positive for HIV. However, this is far away from the distribution in the original data and hence implies serious privacy risks,

[193] Ibid., pp. 13 and 16.
[194] Ibid., p. 16.
[195] Ibid.
[196] Ninghui Li, Tiancheng Li & Suresh Venkatasubramanian, *t-Closeness: Privacy Beyond k-Anonymity and l-Diversity,* IEEE 23rd International Conference on Data Engineering, ICDE, pp. 106-115, 2007, p. 108.
[197] Ibid.

since an individual in this 2-diverse dataset can be considered to be 50% HIV positive, compared to 1% in the original dataset.[198]

Secondly, *l*-diversity provides a rather bad privacy protection when the values to a sensitive attribute in an equivalence class are diverse but semantically similar.[199] For instance, if the sensitive attribute is disease and all values to that attribute in an equivalence class are stomach diseases, then an adversary, knowing that the targeted individual belongs to that equivalence class, can conclude that the data subject has some type of stomach disease.[200] Even if the targeted individual's precise value cannot be identified it still implies privacy risks. The question is however, if the privacy risks discussed in this and the previous paragraph are taken into consideration when it is being determined whether a dataset contains personal data or anonymous data according to the DPD and the GDPR. This very relevant question will be discussed further in section 3.3.

To summarize, *l*-diversity, in the same way as *k*-anonymity, prevents an adversary from singling out an individual. *l*-Diversity further extends *k*-anonymity since it eliminates the possibility for an adversary to obtain sensitive information about a data subject due to the lack of diversity in attributes. However, *l*-diversity cannot guarantee that an adversary never can infer information about an individual in a dataset, since it is plausible that the adversary has all the background knowledge needed in order to exclude all values but one to a certain attribute. Nevertheless, the actor who is trying to render the data anonymous can calculate how hard it would be for an adversary to determine the right value of the targeted individual. If the value of *l* is sufficiently high, it might be considered to be reasonably impossible to link the data to a specific individual. It should be noted that the value of *l* is crucial for whether the technique can provide the privacy guarantees needed in order for the dataset to be regarded as anonymous.[201] Moreover, a dataset that is *l*-diverse can still imply privacy risks if the dataset contains one sensitive attribute that only has two possible values, which representation in the overall distribution are very different. An *l*-diverse dataset can further threaten people's privacy if the attribute values in an equivalence class are diverse but semantically similar. However, it can be questioned whether these privacy threats could lead to that a dataset is considered to contain personal data and thus whether such risks need to be eliminated in order to reach the requisite level of anonymization stated in Recital 26 in the DPD and the GDPR.

[198] See ibid., where this is called a skewness attack.
[199] Ibid., pp. 108-109.
[200] Ibid.
[201] See A29WP (2014), supra note 100, pp. 18-19.

Hence, it can be concluded that not even this technique can a big data actor employ and be certain of that the dataset falls outside the scope of the DPD and in the future the GDPR.

3.2.2.3 t-Closeness

Although *l*-diversity is an improvement of *k*-anonymity it has, as discussed above, some shortcomings. *t*-Closeness is a refinement of *l*-diversity, which aims to overcome these shortcomings by requiring that the distribution of a sensitive attribute in any equivalence class is similar to the initial distribution of the attribute in the original dataset.[202] In other words, "an equivalence class is said to have *t*-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold *t*.[203] Hence, by employing *t*-closeness one can prevent privacy risks like those discussed above, that a person as being part of a certain equivalence class is perceived to have a 50% risk of being HIV positive, when the risk based on the whole dataset is merely 1%. Further, it also prevents the situation where all the values to a sensitive attribute within an equivalence class are diverse but still very similar, which enables an adversary to withdraw sensitive information about an individual (e.g. that the individual has some type of stomach disease), when this distribution does not correspond to the distribution in the entire dataset.

It can be concluded that *t*-closeness further limits the amount of information an adversary can obtain about a specific individual from a dataset, while it still preserves the overall distribution in the original dataset.[204] Hence, this technique can be very useful when it is important that the dataset as a whole is as similar as possible to the original dataset.[205] However, *t*-closeness, as all other anonymization techniques, implies challenges. First of all, it can be difficult to achieve a *t*-close dataset, due to the fact that another constraint is imposed on the data.[206] The dataset should not only be altered so that each equivalence class contains at least *k* records and at least *l* well represented values for each sensitive attribute, but also so that each value in every equivalence class is represented neither more nor less than it is in the entire dataset. Secondly, it can be problematic to measure the distance between the distribution of a sensitive attribute in an equivalence class and the distribution of the attribute in the whole

[202] N. Li et al., supra note 196, p. 106. Ninghui Li, Tiancheng Li and Suresh Venkatasubramanian originated the notion *t*-closeness.
[203] Ibid., p. 109.
[204] Ibid., p. 107.
[205] A29WP (2014), supra note 100, p. 18.
[206] Ibid.

dataset.[207] If that distance cannot be measured properly, the aim of the technique will not be fulfilled. Furthermore, it can be questioned whether *t*-closeness is necessary to achieve in order to reach the requisite level of anonymization stated in Recital 26 in the DPD and the GDPR. What is required in order to meet this standard will be discussed in the following section.

3.3 Evaluation of Anonymization as a Method to Utilize the Benefits of Big Data

As pointed out in the very beginning of this thesis, striking an adequate balance between beneficial uses of big data and privacy risks has been called "the biggest public policy challenge of our time".[208] In a society where our personal data is collected, stored, analyzed and shared daily, a strong protection for our privacy is necessary. However, privacy guarantees implies decreased opportunities to utilize the data for big data analytics. The question is hence whether utilization of personal data is completely incompatible with strong privacy legislation or if there is a solution to this conflict.

First of all, it can be concluded that anonymization, at least in theory, is a method that can address this problem, since if a dataset is rendered anonymous it falls outside the scope of the legislation and thus can be processed freely. In this case it is possible to reap the benefits of big data, while still having a strong protection for our privacy. However, it is shown from the above analysis that a truly anonymized dataset is difficult to achieve and no available anonymization technique can fully be relied on for the aim of enabling broader processing. Both noise addition and permutation are proven to be insufficient if applied alone.[209] Although differential privacy strikes a rather reasonable balance between privacy and utility, it has limitations. An adversary can by posing multiple questions disclose personal information. Thus, queries and answers have to be tracked in order for the technique to be robust against privacy breaches, which makes the technique burdensome, costly and time consuming for the data controller.[210] Moreover, the differential privacy mechanism developed by Dwork implies that the data controller has access to the original data. If Recital 26 in the DPD and the GDPR is interpreted according to the absolute approach, the original dataset must be erased if the data is to be rendered anonymous. It might affect the usefulness of the technique if the controller

[207] N. Li et al. proposes that the Earth Mover Distance (EMD) should be used to measure this distance. However, the EMD measure has several limitations. For more information regarding EMD, how it is used in *t*-closeness, and what the limitations are, see N. Li et al., supra note 196, pp. 110-115.
[208] Tene & Polonetsky (2013), *Privacy and Big Data: Making Ends Meet*, supra note 10.
[209] See A29WP (2014), supra note 100, pp. 13-14.
[210] See Ohm (2010), supra note 114, p. 1756.

will have to base the answers on an already altered version of the data. However, if the relative approach is applied, the data can be considered anonymous in relation to third parties even if the controller retains a copy of the original data. Hence, we will have to await further developments in case law before it can be determined to what extent Dwork's differential privacy mechanism can render data anonymous according to EU data protection legislation. Furthermore, differential privacy can only be employed for the release of data and thus is not sufficient for all purposes. What regards *k*-anonymity, it is proven that merely grouping at least *k* records together does not prevent an adversary from inferring personal information. *l*-Diversity resolves the problem with *k*-anonymity, but unfortunately has other shortcomings. The technique cannot guarantee that an adversary with arbitrary amounts of background knowledge never will be able to retain personal information about an individual. In addition, *l*-diversity is not robust against certain privacy threats if the dataset contains one attribute that only has two possible values, which representation in the overall distribution are very different or if the values within an equivalence class are diverse but semantically similar. *t*-Closeness addresses these shortcomings, but it might, however, be difficult to measure whether a *t*-close dataset has been achieved. To conclude, no of the presented techniques provides a watertight solution. Hence, it can be questioned whether we should continue to put our trust in anonymization for the aim of enabling the society to fully take advantage of the opportunities with big data.

As a matter of fact, this question has been discussed among legal scholars ever since computer scientists, for over a decade ago, proved that anonymized data could be reidentified and thus that anonymization was not as safe as previously thought.[211] Since this revelation several computer scientists have tried to develop new anonymization techniques and refine existing techniques with the aim of providing one technique that is perfectly safe and robust against any sort of attacks.[212] Although progress has been made, the reidentification threat still remains.[213] This has led to a polarization among experts in the field.[214] On one hand, there are scholars like Paul Ohm who is aiming to convince regulators, business actors and the rest of the society that anonymization has failed, that we should abandon our faith in this method since no available technique can achieve a truly anonymized dataset and that we must look for

[211] Rubinstein & Hartzog, supra note 118, p. 709. Ohm (2010), supra note 114, pp. 1706-1707.
[212] See Section 3.2.
[213] Paul Ohm rejects techniques such as differential privacy, since such techniques cannot promise perfect privacy and additionally are more difficult to achieve. See Ohm (2010), supra note 114, pp. 1728, 1755-1757.
[214] Rubinstein & Hartzog, supra note 118, p. 709.

alternative solutions to restore the balance between privacy and utility.[215] On the other hand, there are scholars like Jane Yakowitz, Ann Cavoukian and Khaled El Emam who argue that the likelihood of reidentification for most datasets is, in fact, very low and thus that established anonymization techniques remain sufficient for protecting data subjects and enabling big data analytics.[216]

In order to determine whether existing anonymization techniques are sufficient, one must first answer the question what the techniques should be sufficient for. If the goal is to achieve a truly anonymized dataset that is robust against all types of attacks, one can from the analysis above conclude that every established technique has failed.[217] However, the legislation (i.e. the DPD and the GDPR) does not require 100 % anonymity. The criterion, which an anonymization technique has to live up to, is that it should be impossible to identify an individual, directly or indirectly, using all the means reasonably likely to be used by either the controller or another person.[218] Although 100 % anonymity is the most desirable position, this is not the criterion entrenched in the law.[219] Hence, even if established anonymization techniques are not completely robust, they might still reach the required level of anonymization stated in the law and will thus fulfill their function of enabling broader processing. Accordingly, some techniques, even though they do have limitations, can be of immense value for certain actors.

However, the problem is that it is very difficult to determine when the criterion entrenched in the law has been met. It is clear that perfect anonymity is not needed. The crucial question is rather when data is considered to be personal verses non-personal. The distinction between these two categories has become more and more fluent.[220] What before were seen as deidentified data and thus non-personal, can today be reidentified and will hence be considered personal.[221] To recall, we do know that data is considered personal when the information concern an identified or identifiable natural person, and that an identifiable natural person is someone that can be identified either directly or indirectly using all the means reasonably likely to be used by the controller or another person.[222] We further know that all other data is considered

[215] See Ohm (2010), supra note 114.
[216] See Jane Yakowitz, *Tragedy of the Data Commons,* 25 Harvard Journal of Law & Technology 1, Fall 2011; and Ann Cavoukian & Khaled El Emam, *Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy,* Information & Privacy Commissioner of Ontario, June 2011.
[217] See also Ohm (2010), supra note 114, pp. 1717, 1731 and 1757.
[218] Recital 26 in the DPD and the GDPR.
[219] For a similar reasoning regarding the requirement in the British Data Protection Act see ICO (2012), supra note 95, p. 6.
[220] Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information,* 86 N.Y.U. L. Rev. 1814, 2011, p. 1814.
[221] Ibid.
[222] Recital 26 in the DPD and GDPR.

non-personal. However, it is not an easy task to determine which category certain data falls within based on these vague definitions. As a matter of fact, these definitions have to be vague, since the line between personal data and non-personal data is not fixed, but rather changes in different contexts and over time.[223] Accordingly, it is nearly impossible for any actor to beforehand determine which means that will be considered as reasonably likely to be used in order to reidentify the data, since the capability to reidentify varies due to technological improvements. Further, the fact that the line between personal and non-personal data is so vague, makes it hard to determine what the law count as a reidentification, i.e. what the law count as a privacy breach, and hence what an anonymization technique should prevent.[224] This uncertainty makes it particularly difficult for an actor to determine whether an anonymization technique is sufficient or not for exempting data from the data protection legislation.

From the definition of personal data it is clear that identity disclosure, i.e. when an individual can be singled out from a dataset, count as a reidentification and thus as a privacy breach.[225] It is further clear that attribute disclosure, i.e. when a certain value of an attribute can be linked or inferred to a specific data subject, can be seen as a reidentification.[226] However, it is unclear whether an anonymization technique has to ensure that no attribute disclosures are possible. For instance, as presented above *l*-diversity proves to be robust against extensive attacks, including both identity disclosure and attribute disclosure, but it cannot ensure that an adversary with arbitrary background knowledge never will be able to infer a value of an attribute to a certain data subject. However, if the value of *l* is very high, it might be considered reasonably impossible to identify an individual directly or indirectly, taken into account the time, cost and technology a reidentification would require.[227] The problem is that an actor who is processing the data will be unable to foresee, with any degree of certainty, whether the technique will make it reasonably impossible to identify a data subject, since this criterion is not fixed and hence does not clarify which privacy threats a technique must be robust against.

[223] Schwartz & Solove, supra note 220, p. 1818.
[224] See Felix Wu's discussion regarding that it is crucial to define what counts as a privacy breach to be able to determine what legal and technical tools that are appropriate for striking a balance between privacy and utility. Wu, supra note 154, pp. 1146-1147.
[225] In this case the individual will be regarded as directly identified. See the definition of personal data in article 4(1) in the GDPR and article 2(a) in the DPD.
[226] In this case the individual is seen as identified indirectly. See article 4(1) in the GDPR and article 2(a) in the DPD, as well as CJEU, Case C-582/14, Breyer, para 41.
[227] See Recital 26 in the GDPR.

Moreover, it is further unclear if an adversary's procurement of partial, or uncertain, information counts as a reidentification and a privacy breach according to the law.[228] For example, does the fact that an adversary from an *l*-diverse dataset draws the conclusion that an individual in a certain equivalence class has 50 % possibility of being HIV positive count as a privacy breach according to the law and thus imply that *l*-diversity is insufficient as a technique for avoiding application of the legislation? It is clearly a privacy threat, but it does not imply identification, neither directly nor indirectly, of an individual. From this, one could conclude that the law does not count this as a privacy breach and that even if *l*-diversity does not prevent such privacy threats the technique is still sufficient for rendering personal data non-personal according to the DPD and the GDPR. Thus, *l*-diversity is an example of a technique that, despite its flaws, may be sufficient for reaching the required level of anonymization stated in the law and hence can be of significant value for big data actors. However, an actor can never be certain that *l*-diversity or any other anonymization technique will be sufficient for rendering personal data non-personal, since the concept that defines the boundaries of privacy law is not fixed and the law does not specify what count as a privacy breach.[229]

To conclude, although certain anonymization techniques have the capability of rendering a dataset anonymous according to the law, there is always a risk of reidentification, since as mentioned above this risk is inherent to anonymization. Obviously neither business actors, regulators nor the society at large wants to be stranded with techniques that imply a great risk. This risk both limits innovation and creates unease among people. Hence, there is an urgent need for a new solution to the challenge of striking an adequate balance between privacy and utility, which can offer both big data actors and individuals a larger degree of certainty.

4. ALTERNATIVE SOLUTIONS

As concluded above, it is nearly impossible for actors to know beforehand if a particular anonymization technique is sufficient for rendering personal data non-personal, since the line between these two categories of data is not fixed. However, what actors do know is that it is much more difficult to anonymize data than originally thought. Computer scientists have shown that what before were characterized as non-personal data can today actually be linked

[228] See Wu, supra note 154, pp. 1159-1165.
[229] See Schwartz & Solove, supra note 220, p. 1818; and Wu, supra note 154, part III.A.

to individuals,[230] and deidentified data can often be reidentified (e.g. as in the Netflix Prize example).[231] Hence, more data than was originally intended is now falling under the category of personal data.[232] Accordingly, it is more difficult than ever to make identification reasonably impossible and hence to reach the required level of anonymization stated in the law. Thus it is harder than before to fall outside the scope of the legislation. This further implies that it is significantly more difficult to utilize the benefits of big data. Thus, it is crucial to restore the balance between privacy and utility in the law.[233] How this should be done is, however, not an easy question.

4.1 Abandon Anonymization and the Entire Concept of Personal Data

Ever since computer scientists proved that anonymized data could be reidentified, legal scholars have tried to find a solution to restore the balance between privacy and utility in data protection legislation. Almost every single information privacy statute in the world, including the DPD and the GDPR, rests on the distinction between personal and non-personal data.[234] As concluded above, this distinction defines the boundaries of such legislation. Personal data falls under the scope of the legislation, whereas non-personal data falls outside the scope of the legislation.[235] The intent with the DPD and the GDPR is that both these statutes should have such boundaries and thus not cover all data. The legislator even recognizes a possibility to exempt data originally personally identifiable from the legislation through anonymization.[236] This recognition is intended to create a balance between privacy and utility by enabling uses of information derived from individuals as long as those individuals are not directly or indirectly identifiable. However, the revelation that anonymized data often can be reidentified has disrupted this balance.[237] Hence, Paul Ohm argues that we must abandon our faith in anonymization and find a new solution.[238] According to Ohm the solution is neither to wait

[230] For example, before IP-addresses were classified as non-personal data, but today it is seen as personal information in some cases and thus falls under the scope of the EU data protection legislation. See CJEU, Case C-582/14, Breyer.
[231] Schwartz & Solove, supra note 220, p. 1814.
[232] Paul Ohm argues that the fact that anonymized data can easily be reidentified has made laws like the DPD overbroad. He refers to the DPD as a law that was meant to have limits but which has now been rendered limitless. See Ohm (2010), supra note 114, p. 1741. Paul Schwartz and Daniel Solove also notice that too much information is considered to be personal data according to the EU data protection legislation. See Schwartz & Solove, supra note 220, p. 1871.
[233] Paul Ohm also recognizes this imbalance. He argues that ”the loss of robust anonymization reveals the lurking imbalance in … privacy laws” and states that the DPD has come to protect privacy too much. Ohm (2010), supra note 114, pp. 1740-1741.
[234] Schwartz & Solove, supra note 220, p. 1816.
[235] Ibid. See also Section 3.1 above.
[236] Recital 26 in the DPD and the GDPR.
[237] Ohm (2010), supra note 114, pp. 1740-1741.
[238] Ibid., pp. 1742-1743.

for technologists to develop new privacy-protecting techniques nor to implement a legislative ban on reidentification.[239] Ohm argues that the problem does not lie in the failure of anonymization alone, but essentially lies in the mere concept of personal data.[240] As concluded above, personal data is not a fixed concept but changes as technology develops. Ohm means that personal data is an ever-expanding category, which will never stop growing until it includes all data.[241] This entails that personal data, as a distinguishing concept of what falls within and outside the scope of the legislation, can never strike an adequate balance between privacy and utility, since the category of personal data will expand until everything falls within the legislation and thus no room is left for utility.

Therefore, Ohm proposes that we should abandon the entire concept of personal data and find a new organizing principle.[242] Ohm suggests that we instead of distinguishing between personal and non-personal data should distinguish between different types of database owners and different types of databases.[243] We should, by considering a series of factors, identify situations in which harm outweighs the benefits of free information flow and create sector-specific legislation that targets such situations.[244] Although Ohm may be right that personal data is not an optimal concept on which to regulate information privacy, an organizing principle that can replace personal data has not yet been developed. Less than a year ago all member states within the EU agreed to adopt a regulation that comes into force in May 2018. The scope of this new regulation merely relies on the distinction between personal and non-personal data. That all member states would agree to change the core concept in the regulation is rather unlikely. Hence, although regulators and legal scholars should be encouraged to develop a new organizing principle for the future, less disruptive measures can be taken in the meantime, which would significantly decrease the current imbalance between privacy and utility, and thus would make it less difficult for the society to utilize the benefits of big data.

[239] Ohm means that no new privacy-protecting technique can accomplish what was once promised by anonymization and therefore we should not wait for technology to us. He further argues that banning reidentification is not a suitable solution since it will be impossible to enforce. See Ohm (2010), supra note 114, p. 1745.
[240] Ibid., p. 1742.
[241] Ibid.
[242] Ibid., p. 1743. To clarify, abandoning personal data as a concept means abandoning anonymization as well, since anonymization is a process conducted for the aim of rendering personal data non-personal. In other words, if anonymization is to be of any value for big data actors and if there is to be any incentive to anonymize, the law has to rely upon the concept of personal data (i.e. the boundaries of the law has to be determined based on whether data is personal or not).
[243] Ibid., p. 1759.
[244] Ibid., pp. 1759-1769.

4.2 Retain the Concept of Personal Data and Anonymization but Establish Clarifying Standards

As briefly touched upon above, measures can be taken to restore the intended boundaries of the data protection legislation within the EU, without abandoning the concept of personal data. Paul Schwartz and Daniel Solove even argue in their path-breaking article *The PII Problem: Privacy and a New Concept of Personally Identifiable Information* that information privacy law needs a concept of personal data.[245] Based on the analysis above, that it would certainly be difficult to achieve an agreement between all member states of the EU to change the newly adopted regulation, it can be concluded that we, at least for now, need to retain the concept of personal data.[246] However, clearly something needs to be done, because currently the legislation does not strike an adequate balance between privacy and utility.

Even if we have concluded that we cannot abandon personal data as a concept yet, the question still remains whether we should continue to put our trust in anonymization.[247] The facts that a risk of reidentification is inherent to anonymization and that no established anonymization technique can be fully relied upon, suggests that we should abandon anonymization. However, this might not be needed. The risk of reidentification is not equally large in all situations and as discussed above some anonymization techniques may in certain situations be sufficient for living up to the required level of anonymization entrenched in the law. Hence, it seems foolish to dismiss anonymization completely, when it is proven to have the capability of being sufficient for the purpose of enabling broader processing.

Nevertheless, it has become easier to reidentify data and thus more difficult to anonymize data, which has disrupted the balance between privacy and utility in the law. The balance needs to be restored, which can be done by establishing standards on what is required in order to fall outside the scope of the legislation. The establishment of such standards does not require any amendment of the regulation, which is why this solution is particularly appropriate. How these standards should be shaped, what they should cover and who should enact them, will be elaborated on in the following three paragraphs.

The content of these standards should depend on the level of risk for reidentification. As noted above the risk for reidentification is not equally large in all situations, but depends to a

[245] Schwartz & Solove, supra note 220, p. 1817.
[246] See Section 4.1.
[247] As highlighted in supra note 242, abandoning personal data as a concept implies abandoning anonymization as well. However, it is not necessary the same the other way around. Retaining personal data as a distinguishing principle of what falls under the scope of the law does not automatically imply that anonymization has to be retained as a solution for enabling broader research.

great extent upon the specific context.[248] Hence, the measures that need to be taken in order to fall outside the scope of the legislation vary depending on the circumstances surrounding the processing of the data. Thus, at least the following factors should be taken into consideration when establishing standards. First of all, the volume of the data can affect the risk of reidentification.[249] Large datasets often have a higher degree of unicity and therefore is more reidentifiable than small datasets.[250] This is evidenced by the fact that all famous, successful reidentification attempts have included large datasets.[251] Hence, the larger a dataset is, the more is required from the controller in order to fall outside the scope of the DPD and the GDPR, which should be embraced when developing standards. Moreover, the character of the data also affects the reidentification risk. Sensitive data, like health information for example, is generally more likely to be targeted by attackers and accordingly the risk of reidentification is greater.[252] In addition, the leakage of sensitive data, as opposed to non-sensitive data, can cause significantly more harm to individuals.[253] For instance, it is a greater intrusion in someone's privacy if it is being revealed that an individual is HIV positive, than if it is being revealed that a certain individual buys four liters of milk every week. Hence, the more sensitive the data is, the more important is it that the data remains deidentified and thus that data controllers take appropriate safety measures. Further, the potential harm has to be weighed against the utility of the data.[254] If data is particularly useful for society, it might be reasonable to lower the requirements for falling outside the scope of the legislation and in that way stimulate important data uses.[255] Furthermore, the risk for reidentification naturally varies depending on what the data is used for and who is having access to the data.[256] If the data is solely going to be used for internal purposes, such as improving an actor's product, website or customer service, and thus no one outside of the organization will get access to the data, the reidentification risk is quite small. If the data is released to a trusted party, such as a universi-

[248] See Schwartz & Solove, supra note 220, p. 1847.
[249] Rubinstein & Hartzog, supra note 118, p. 741.
[250] "Unicity quantifies how much outside information one would need, on average, to reidentify a specific and known user in an … anonymized data set. The higher the data set's unicity is, the more reidentifiable it is". Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh & Alex Pentland, *Unique on the Shopping Mall: On the Reidentifiability of Credit Card Metadata,* 347 Science 536, 2015, p. 537.
[251] Ohm (2010), supra note 114, p. 1766. See for example the Netflix Prize discussed in Section 3.2.1.1.
[252] Rubinstein & Hartzog, supra note 118, p. 741.
[253] Ibid., pp. 735 and 741.
[254] Ibid., p. 735.
[255] See ibid, pp. 742-743; and Yves-Alexandre de Montjoye, Jake Kendall & Cameron F. Kerry, *Enabling Humanitarian Use of Mobile Phone Data*, Issues in Technology Innovation, Brookings Center for Technology Innovation, 2014, p. 4. This is in line with that Union or Member State law may provide derogations from certain provisions in the GDPR for scientific purposes or purposes in the public interest. See article 89 in the GDPR.
[256] Rubinstein & Hartzog, supra note 118, pp. 741-742.

ty, for research, the risk is a bit greater, but not as great as if the dataset is released to the general public. Also these variations in risk should be taken into account when drawing up standards on what is required in order to avoid application of the law. What is further important is to specify the nature of potential adversaries. Depending on whom the data is released to, what it is used for and what type of data it is, the characteristics of potential adversaries will vary. How protective measures that need to be taken to exempt data from the legislation should be based on the motivation for and the likelihood that an adversary will attempt to reidentify the data, as well as the abilities of potential adversaries in terms of for example computational power.[257] The nature of the data controller should also be considered. Such as how much effort a controller can spend on performing a deidentification process and other safety measures.[258] All the factors highlighted in this paragraph should be considered when establishing standards for what is required in order to make identification reasonably impossible and thus to avoid application of the legislation.

Further, what all these factors indicate is that all data cannot be subject to the same requirements, since the reidentification risk varies significantly. To require the same level of protection for all data would be a waste of useful innovation. Clearly, one size does not fit all.[259] Instead, the requirements for falling outside the scope of the legislation should be dependent upon context. However, information privacy law cannot be too specified, since it quickly becomes outdated due to the fast technological development and since laws, especially EU regulations, are difficult to amend.[260] Hence, the better alternative to regulatory specificity is to establish standards, which can be changed and improved regularly based on the input from different experts and stakeholders.[261]

These standards should first of all include guidance on how certain data should be deidentified. In other words, the standards should clearly set out which anonymization technique to apply based on specific circumstances. However, the standards should not only include guidance on which anonymization technique to apply, but also guidance on which other safety measures that need to be taken in order to make it reasonably impossible to identify any individual in a certain dataset. Such safety measures can be different types of data controls,

[257] Ibid., p. 735. See also Wu, supra note 154, p. 1148.
[258] Rubinstein & Hartzog, supra note 118, p. 735. For instance, small and medium-sized enterprises may not have the same capacity to perform and test different safety measures. Hence, it is reasonable to require less from such actors, in order to enable data analysis conducted by smaller actors and to promote fair competition.
[259] Ibid., p. 736.
[260] Schwartz & Solove, supra note 220, pp. 1871-1872.
[261] Ibid. See also Alexandra Wood, Edo Airoldi, Micah Altman, Yves-Alexandre de Montjoye, Urs Gasser, David O'Brien & Salil Vadhan, *Comments on the Proposed Rules to Revise the Federal Policy for the Protection of Human Subjects,* Harvard University Privacy Tools Project, 2016, p. 14.

like public warrants to not reidentify, contracts prohibiting reidentification, audit trails and access limitations. A few examples can be taken to illustrate what different standards could include. For instance, a standard targeting large datasets that are going to be analyzed for internal purposes only, may require that the actor applies the technique of *l*-diversity to deidentify the data as well as publicly commits not to try to reidentify the data.[262] Moreover, data controllers that are going to release a dataset to one or a few trusted parties, may be required to, except rendering the dataset *l*-diverse and publicly commit to not reidentify the data, contractually prohibit also downstream users from trying to reidentify the data.[263] If the dataset contains particularly sensitive information, such as health information, the controller may be required to insert audit trails as well, which can track how the data is being used.[264] The standard could also require that the initial data controller impose sanctions, such as fines, on downstream recipients who violate a contractual obligation to not try to reidentify a dataset. For the enforcement of such sanctions, audit trails can be of immense value. If the recipients of the data are less trustworthy, the standard could even require that access controls are implemented, which limit the ways the recipients can interact with the data.[265] Where suitable, instead of access controls, data controllers could be required to use interactive deidentification techniques, such as differential privacy, so that only limited parts of a dataset is being released. Standards targeting public releases of large datasets should include the most extensive requirements, since the reidentification risk is greatest in these situations. Such standards could even include a presumption that publicly released datasets are identifiable and hence falls under the scope of the legislation, unless a data controller fulfills all requirements in the standard.[266] As highlighted above, this paragraph has only presented a few examples of what standards could include. To formulate appropriate standards that could actually be adopted would require research that lies beyond the aim of this thesis.

Nevertheless, a discussion is needed regarding who should have the responsibility to establish these standards. When the DPD was adopted an organ called the A29WP was set up.[267] The A29WP has throughout the years issued opinions on questions of data protection.[268] These opinions have provided useful guidance on how to interpret the DPD. The A29WP will be

[262] The US Federal Trade Commission also suggests such public commitment. See FTC Report, *Protecting Consumer Privacy in an Era of Rapid Change,* Recommendations for Businesses and Policymakers, 2012, p. 22.
[263] Ibid.
[264] Audit trail is a type of software that tracks usage in data. See for example Ohm (2010), supra note 114, p. 1756.
[265] Ibid.
[266] See Rubinstein & Hartzog, supra note 118, p. 755.
[267] Article 29 in the DPD.
[268] The tasks of the A29WP are stated in article 30 in the DPD.

replaced by a new body, the European Data Protection Board (EDPB), when the GDPR enters into force in May 2018.[269] One of the EDPB's tasks is to issue guidelines, recommendations and best practices in order to encourage consistent application of the GDPR.[270] Hence, it is natural to suggest that standards or guidelines, on what is required in order to fall outside the scope of the regulation, should be established by the EDPB. Until the EDPB has been established, the A29WP could preferably start developing the standards.

To have clear standards or guidelines to follow would clearly be useful for actors wishing to reap the benefits of big data. However, if standards are going to improve the situation for such actors, they will have to provide legal certainty. Legal certainty could be created through the introduction of guarantees. Such guarantees would then imply that a certain dataset is guaranteed to fall outside the scope of the GDPR if the actor, at all times,[271] meets every single requirement stated in a standard targeting the relevant dataset. This would certainly make it easier to conduct big data analytics, since the establishment of such safe harbors would enable actors to beforehand determine which measures that are necessary to take in order to avoid application of the law. Hence, the implementation of standards working as guarantees would reintroduce boundaries to the law. The scope of the law would still be more extensive than originally intended, since the revelation of the reidentification risk has made it more difficult to anonymize data and thus more data than previously thought is now falling under the regulation. However, even if more data counts as personal data nowadays, the introduction of such standards presented in this section would at least create a line between personal and non-personal data, which actors could adhere to. This would make it easier as well as less risky to utilize data originated from individuals in big data analytics and would hence improve the balance between privacy and utility in the law. Moreover, the introduction of guarantees would also place the outer responsibility for developing new, sufficient deidentification techniques and data controls on a EU organ rather than on single actors, which might seem more reasonable. Further, creating safe harbors will encourage actors to anonymize and take additional safety measures to protect people's privacy. If the situation remains unchanged, actors will stop anonymizing data, since it is not economically justifiable to invest time and money on anonymization if it cannot be guaranteed that the data falls outside the scope of the regulation. This will result in larger danger to privacy than the reestablishment of legal boundaries

[269] See Recital 139 in the GDPR.
[270] Article 70(1)(e) in the GDPR.
[271] This means that an actor has to constantly control that all requirements are fulfilled, which include keeping track of any changes in the relevant standard and update the protection if needed.

would. Hence, it is crucial to restore the room for utility, both for the development of our society and for the protection of our privacy.

To summarize, the revelation that deidentified data can be reidentified has made it more difficult than ever to fall outside the scope of the law, and has thus disrupted the intended balance between privacy and utility in European data protection legislation. This balance can be restored by establishing clarifying standards on what is required in order to avoid application of the GDPR.[272] The requirements stated in these standards should be dependent upon context and should be more or less extensive based on the reidentification risk. The standards should not only state how particular data should be anonymized, but also what other safety measures that needs to be taken in order to make it reasonably impossible to identify an individual in a dataset. Such safety measures could include public warrants to not try to reidentify, contracts prohibiting reidentification, audit trails and access limitations. The standards should preferably be enacted by the EDPB when the GDPR has entered into force. Further, if an actor fulfills every single requirement in a standard targeting the relevant dataset, it should be guaranteed that the dataset falls outside the scope of the regulation. The establishment of such safe harbors would reintroduce boundaries to the law. The reintroduction of boundaries would in turn decrease the current imbalance between privacy and utility in the GDPR, and would thus make it easier to utilize the benefits of big data.

Lastly, it should be noted that there might be other solutions than those discussed herein to the challenge of restoring the balance between privacy and utility in data protection law. However, to abandon the concept of personal data and establish a new organizing principle or to retain the concept of personal data but introduce contextually sensitive rules, have been the most discussed approaches within scholarly works. As concluded above, the latter approach is the more suitable one under the given circumstances, since to change the core concept in the GDPR would be practically impossible.

5. CONCLUDING REMARKS

In today's society, where every step we take is recorded, stored, analyzed and shared, far-reaching data protection legislation is necessary. However, it is nearly impossible to comply with the EU data protection rules while conducting big data analytics. As presented in the very beginning of this thesis, big data can provide immense value for humankind. Hence, EU

[272] Since the drafting of standards will require more research and hence some time, it is suitable that these standards only refers to the application of the GDPR, and not the DPD.

data protection legislation has to, except fulfilling its primary goal of protecting every one's right to privacy, at least make it feasible to utilize the benefits of big data. Before, when it was still possible to accomplish bulletproof anonymization, the DPD actually achieved striking an adequate balance between privacy and utility. However, the revelation that anonymized data can be reidentified has disrupted that balance. Currently, it is harder than ever to fall outside the scope of the legislation and hence it is equally hard to reap the benefits of big data. Since the GDPR relies on the same distinction between personal and non-personal data as the DPD, the balance between privacy and utility is lost in a statute that has not even entered into force. Clearly, something needs to be done in order to prevent that a boundless regulation comes into force in 2018. Above has been suggested that the current imbalance between privacy and utility can, at least, be decreased through the establishment of clarifying standards. Such standards can provide legal certainty to actors by guaranteeing that a specific dataset will fall outside the scope of the GDPR if all requirements stated in a standard targeting the relevant dataset are fulfilled. Although the conclusion of this thesis is that the establishment of clarifying standards is the best alternative for restoring the balance between privacy and utility in the law, further research is necessary in order to form a solution that can actually be implemented. Hence, scholars and policymakers are hereby encouraged to continue the work of finding a solution to what has been called 'the biggest public policy challenge of our time' and to again make it possible to utilize the benefits of big data.

BIBLIOGRAPHY

Legislative Acts

Charter of Fundamental Rights of the European Union OJ C 326/391, 26.10.2012

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

European Convention on Human Rights

Regulation 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

Court Cases

Court of Justice of the European Union, Judgment of 19 October 2016, Breyer, C-582/14, EU:C:2016:779

Court of Justice of the European Union, Judgment of 1 October 2015, Bara and Others, C-201/14, EU:C:2015:638

Court of Justice of the European Union, Judgment of 13 May 2014, Google Spain, C-131/12, EU:C:2014:317

Court of Justice of the European Union, Judgment of 8 April 2014, Digital Rights Ireland, C-293/12 and C-594/12, EU:C:2014:238

Court of Justice of the European Union, Judgment of 17 October 2013, Schwarz, C-291/12, EU:C:2013:670

Court of Justice of the European Union, Judgment of 30 May 2013, Worten, C-342/12, EU:C:2013:355

Court of Justice of the European Union, Judgment of 6 November 2003, Lindqvist, C-101/01, EU:C:2003:596

Court of Justice of the European Union, Judgment of 20 May 2003, Rechnungshof v Österreichischer Rundfunk and Others, C-465/00, C-138/01 and C-139/01, EU:C:2003:294

Books

Colonna, L., *Legal Implications of Data Mining: Assessing the European Union's Data Protection Principles in the Light of the United States Government's National Intelligence Data Mining Practices*, Ragulka förlag, 2016

Dwork, C., *Differential Privacy,* in *Automata, Languages and Programming*, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II, Springer Berlin Heidelberg, 2006

Kalyvas, J.R., Overly, M.R., *Big Data: A Business and Legal Guide*, Auerbach Publications, 2014

Mayer-Schönberger, V., Cukier, K., *Big Data: A Revolution That Will Transform How We Live, Work and Think*, John Murray, 2013

Mohanty, H., Bhuyan, P., Chenthati, D., *Big Data: A Primer,* Studies in Big Data, Vol. 11, Springer India, 2015

Rouvroy, A., Poullet, Y., *The Right to Informational Self-Determination and the Value of Self-Development: Reassessing the Importance of Privacy for Democracy,* in *Reinventing Data Protection?,* Springer, 2009

Sandgren, C., *Rättsvetenskap för uppsatsförfattare: ämne, material, metod och argumentation*, Norstedts Juridik, 2015

Van Klink, B., Taekema, S., *On the Border: Limits and Possibilities of Interdisciplinary Research*, in *Law and Method*, Tubingen, Mohr Siebeck, 2011

Journal Articles

Aggarwal, C.C., *On k-Anonymity and the Curse of Dimensionality,* Proceedings of the 31st International Conference on Very Large Data Bases (VLDB) 901, 2005

Butler, D., *When Google Got Flu Wrong: US Outbreak Foxes a Leading Web-based Method for Tracking Seasonal Flu,* 494 Nature 155, Macmillan Publishers Limited, 2013

Chin, A., Klinefelter, A., *Differential Privacy as a Response to the Reidentification Threat: the Facebook Advertiser Case Study,* 90 North Carolina Law Review 1417, 2012

De Montjoye, Y-A., Radaelli, L., Singh, V.K., Pentland, A., *Unique on the Shopping Mall: On the Reidentifiability of Credit Card Metadata,* 347 Science 536, 2015

Jungar, E., *Big Data: Mind the Gap – Regulation Meets Reality,* Juridisk Publikation, number 1, 2016

Kemp, R., *Big Data and Data Protection,* White Paper, Kemp IT Law, 2014

Li, N., Li, T., Venkatasubramanian, S., *t-Closeness: Privacy Beyond k-Anonymity and l-Diversity,* IEEE 23rd International Conference on Data Engineering, ICDE, pp. 106-115, 2007

Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M., *l-Diversity: Privacy Beyond k-Anonymity*, ACM Transactions on Knowledge Discovery from Data, Vol. 1, No. 1, Article 3, March 2007

Munir, A.B., Yasin, S.H.M., Muhammad-Sukki, F., *Big Data: Big Challenges to Privacy and Data Protection*, 9 International Journal of Social, Education, Economics and Management Engineering 355, 2015

Narayanan, A., Shmatikov, V., *Robust De-anonymization of Large Sparse Datasets,* 2008 IEEE Symposium on Security and Privacy, pp. 111-125, IEEE, May 2008

Ohm, P., *The Underwhelming Benefits of Big Data,* 161 University of Pennsylvania Law Review 339, 2013

Ohm, P., *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization,* 57 UCLA Law Review 1701, 2010

Rubinstein, I.S., Hartzog, W., *Anonymization and Risk,* 91 Washington Law Review 703, 2016

Rubinstein, I.S., *Big Data: The End of Privacy or a New Beginning?,* International Data Privacy Law, Vol. 3, No. 2, pp. 74-87, 2013

Sweeney, L., *k-Anonymity: A Model for Protecting Privacy,* International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), pp. 557-570, 2002

Sweeney, L., *Achieving k-Anonymity Privacy Protection Using Generalization and Suppression,* International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), pp. 571-588, 2002

Sweeney, L., Samarati, P., *Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression,* IEEE Security and Privacy, 1998

Schwartz, P.M., Solove, D.J., *The PII Problem: Privacy and a New Concept of Personally Identifiable Information,* 86 New York University Law Review 1814, 2011

Schwartz, P.M., *Property, Privacy, and Personal Data*, p. 2056, 117 Harvard Law Review 2055, 2004

Tene, O., Polonetsky, J., *Big Data for All: Privacy and User Control in the Age of Analytics,* 11 Northwestern Journal of Technology and Intellectual Property 239, 2013

Tene, O., Polonetsky, J., *Privacy and Big Data: Making Ends Meet,* 66 Stanford Law Review 25, 2013

Tene, O., Polonetsky, J., *Privacy in the Age of Big Data: A Time for Big Decisions,* 64 Stanford Law Review Online 63, 2012

Wu, F.T., *Defining Privacy and Utility in Data Sets,* 84 University of Colorado Law Review 1117, 2013

Yakowitz, J., *Tragedy of the Data Commons,* 25 Harvard Journal of Law & Technology 1, 2011

Zarsky, T.Z., *Desperately Seeking Solutions: Using Implementation-based Solutions for the Troubles of Information Privacy in the Age of Data Mining and the Internet Society*, 56 Maine Law Review 13, 2004

Zuiderveen Borgesius, F.J., *Singling Out People Without Knowing Their Names – Behavioural Targeting, Pseudonymous Data, and the New Data Protection Regulation*, 32 Computer Law & Security Review 256, 2016

Reports, Surveys, Opinions, Papers and Comments

Article 29 Working Party (A29WP), Opinion 05/2014 on Anonymisation Techniques

Article 29 Working Party (A29WP), Opinion 4/2007 on the Concept of Personal Data

Cavoukian, A., El Emam, K., *Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy,* Information & Privacy Commissioner of Ontario, June 2011

De Montjoye, Y-A., Kendall, J., Kerry, J.F., *Enabling Humanitarian Use of Mobile Phone Data*, Issues in Technology Innovation, Brookings Center for Technology Innovation, 2014

European Commission, *Special Eurobarometer 431 Data Protection Summary,* 2015

FTC Report, *Protecting Consumer Privacy in an Era of Rapid Change,* Recommendations for Businesses and Policymakers, 2012

Information Commissioner's Office (ICO), *Big Data and Data Protection*, 2014

Information Commissioner's Office (ICO), *Anonymisation: Managing Data Protection Risk Code of Practice,* 2012

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, Report – McKinsey Global Institute, 2011

Rubinstein, I.S., *Framing the Discussion*, Brussels Privacy Symposium on Identifiability: Policy and Practical Solutions for Anonymisation and Pseudonymisation, 2016

Wood, A., Airoldi, E., Altman, M., De Montjoye, Y-A., Gasser, U., O'Brien, D., Vadhan, S., *Comments on the Proposed Rules to Revise the Federal Policy for the Protection of Human Subjects,* Harvard University Privacy Tools Project, 2016

Other Sources

Articles from Newspapers and Magazines

Duhigg, C., *How Companies Learn Your Secrets*, The New York Times Magazine, 16 February 2012, http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all, accessed 1 March 2017

Dwork, C., *A Firm Foundation for Private Data Analysis,* 54 Communications of the ACM 86, 2011, available at http://cacm.acm.org/magazines/2011/1/103226-a-firm-foundation-for-private-data-analysis/fulltext

Lazer, D., Kennedy, R., *What We Can Learn From the Epic Failure of Google Flu Trends*, Wired Science, 2015, available at https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

Lohr, S., *The Age of Big Data,* The New York Times, 11 February 2012, http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all, accessed 12 December 2016

Narayanan, A., Shmatikov, V., *Privacy and Security: Myths and Fallacies of 'Personally Identifiable Information'*, 53 Communications of the ACM 24, 2010, available at https://www.cs.utexas.edu/~shmat/shmat_cacm10.pdf

Williams, R., *Web Browsers: A Brief History*, The Telegraph, 2 May 2015, http://www.telegraph.co.uk/technology/microsoft/11577364/Web-browsers-a-brief-history.html, accessed 12 December 2016

Websites

Google Search Statistics, http://www.internetlivestats.com/google-search-statistics/, accessed 24 November 2016

IBM, *Bringing Big Data to the Enterprise: What is Big Data,* https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html, accessed 12 December 2016

Souza, R., Trollinger, R., Kaestner, C., Potere, D., Jamrich, J., *How to Get Started with Big Data,* BCG perspectives by the Boston Consulting Group, 29 May 2013, https://www.bcgperspectives.com/content/articles/it_strategy_retail_how_to_get_started_with_big_data/, accessed 19 November 2016

The Netflix Prize, http://www.netflixprize.com/, accessed 31 January 2017

Wiggins, L., *If Big Data and Analytics Exist in a Silo, Does the Outcome Matter?*, IBM Big Data and Analytics Hub, 25 February 2014, http://www.ibmbigdatahub.com/blog/if-big-data-and-analytics-exist-silo-doesoutcome-matter, accessed 18 November 2016

Dictionaries

Gartner IT Glossary, http://www.gartner.com/it-glossary/big-data, accessed 18 November 2016