UNIVERSITÀ DEGLI STUDI DI PADOVA

Department of Political Science, Law, and International Studies

Master's Degree in European and Global Studies

The right to privacy in a Big Data society. Merits and limits of the GDPR

by

TARLAN ISMAYILOV

No. 1237018

Supervisor: Prof. GUIDO GORGONI

A.Y. 2021/2022

Introduction

Technologies are bursting into all spheres of our lives. Such is the level of change that some of these technologies entail that they have come to be described as disruptive. For example, the phenomenon of the internet, cloud computing, and big data are disruptive technologies that are revolutionizing the way our world works.

And with these technologies, data is becoming the most precious asset. More data is being created today than ever before in history and collecting, storing, and processing it is possible more easily than ever before. Everything from social networks, card purchases, phone calls and so many other everyday gestures generate data, the study of which is a source of incalculable value.

Companies offer free access services in exchange for being able to access our data and use it for a myriad of purposes, many of which are not even known at the time the data is analyzed.

Specifically, big data is the set of technologies that make it possible to process massive amounts of data from disparate sources, with the aim of being able to give them usefulness that provides value. This could be discovering patterns in the behavior of an organization's customers in order to create much more effective targeted advertising, predicting economic trends, or discovering previously unknown relationships between variables that can open the door to innovation.

However, these new big data opportunities are also accompanied by risks. Perhaps one of the most relevant is the risk that this massive data analysis poses to people's privacy. In this paper, we will examine this risk in detail and analyze it from a legal point of view. Technology evolves at such a rapid pace that sometimes the rules are unable to provide a solution to the new problems that arise.

The paper is divided into four chapters. Thus, in Chapter I we will approach the concept of big data, its characteristics and the formidable opportunities it can bring. We will also briefly review some of the risks it creates, specifically (i) the risk of falling into a blind trust in the algorithms that analyze the data, so that no one reviews the conclusions drawn by the machines, leading to automated decision-making; and (ii) the risk of not checking whether the relationships that appear to be found between variables are true or merely random.

The rest of the paper will focus on the risk that big data poses to privacy and data protection. Thus, in Chapter II we will introduce the concept of personal data and the legal

framework of data protection, as well as the impact that big data has on these regulations. In this chapter the main instruments of data protection law are also discussed. We will deal with consent as an instrument that legitimizes that personal data can be collected and processed in accordance with the law.

Chapter III presents anonymization, the mechanism by which data are made anonymous, and which means that they are no longer personal data, so that they can be processed without being subject to the provisions of data protection regulations.

Subsequently, after having analyzed anonymization and pseudonymization big data, we analyze in Chapter V whether the GDPR has been an effective tool in overcoming problems caused by big data and merits and limitations it brings to the data protection sphere.

Chapter 1: The Big Data Revolution

1.1 Big Data: a new paradigm in the data processing

In the last two decades, the Internet has profoundly changed the way companies work, governments, and especially the way people live and communicate. This communication and access to global, immediate, and low-cost information has revolutionized the economy, society, and even politics. If we refer to the "society" or "information age", as defined by Manuel Castells in "The Information Age", this expression refers to that society where the manipulation and management of information have replaced the control and optimization of resources in industrial processes[1].

The information available has multiplied due to three factors, all three linked to the advancement of new information technologies and the Internet: the increase in the speed of information transmission, the increase in the storage capacity, and finally the innovations of physical hardware (computers as data processors). The social actors that intervene in this type of society according to Castells are the users (citizens, companies, public administration, and governments), the technical actors (the technical infrastructure itself: the terminals, the networks, and the servers where all this information is stored), and finally the contents (which can be tangible or intangible, services or infomediation). The information society would affect all fields of socialization and community life: the economy, legislation, training, culture, promotion, and attitudes in general[2]. What characterizes the current technological revolution is not only the centrality of knowledge and information but "the application of this knowledge and information to knowledge generation and information processing/communication devices, in a feedback loop between innovation and its uses"[3].

From its innovation until the '90s, the Internet lived a time free from capitalism, given its absence of monetization, and to a certain extent free of government intervention. The growing monetization of the network begins when it becomes a place of commerce, initially of goods not physically available, at which time the governments of the different states and some international corporations begin to perceive the need for more secure commerce through the Internet, further facilitating the development of network monetization[4]. In addition to the

[1] Castells, Manuel. "The Information Age: Economy, Society and Culture: The Network Society". (1999)
[2] Idem.
[3] Idem.
[4] Idem.

increase in connection speed and storage capacity, two factors have contributed to this monetization process: lower prices and increased accessibility. According to the computer security expert Bruce Schneier, in any society, the exponential development of technology causes the mechanisms of power in general to multiply. In the case of the Internet, initially, it was a type of distributed power that gained territory: the lack of regulation and legislation of the network allowed the emergence of this decentralized power since this new tool provided coordination and efficiency to the masses and cyberactivism. We speak of powers in both a positive and negative sense: grassroots social movements, dissident groups, hackers, cybercriminals, etc. This was possible since the restrictions established by traditional power have not yet adapted to this new context[5]. Power distributed in the network at the beginning seemed invincible; its potential was expressed, for example, in the so-called "Arab Spring" and the fall, among others, of the dictator Mubarak in Egypt, largely thanks to social networks. But little by little, traditional powers have found and established new forms of control over what happens on the Internet: these are organized institutional powers, such as governments or some large international corporations.

According to Schneier, this progressive advance of the traditional powers, in addition to the aforementioned monetization of the network, is due to the birth and expansion of the cloud life of individuals: the set of manifestations of these to the network, which is hosted on servers from private corporations such as Google, Apple, Microsoft or Facebook. In this sense, the limitations imposed by traditional power can be of two types: those that directly affect the freedoms of communication and movement (such as censorship), and on the other hand, what concerns us here: interference with the right to privacy of citizens[6].

If just a few years ago we were talking about the revolution brought about by the Internet, today we are faced with a new phenomenon, a technological trend that is less visible but just as powerful in terms of transformation: big data. While it is true that they are different realities, it is also true that the Internet greatly facilitates the collection and transfer of data on which big data is based.

Big data is a term that refers to the enormous growth in access to and use of automated information. It refers to the gigantic amounts of digital information controlled by companies,

[5] Bruce Schneier. “Liars and Outliers: Enabling the Trust that Society Needs to Thrive and Carry On” Indianapolis: John Wiley & Sons (2012).
[6] Idem.

authorities, and other organizations, which are subject to extensive analysis based on the use of algorithms[7]. It is not a technology in itself, but rather a working approach to obtaining value and benefits as a result of the processing of the large volumes of data that are being generated every day[8].

The main idea is that by processing massive amounts of information, something that has been impossible until now, we can understand things that were previously unknown when we only analyzed small amounts of information, and discover or infer facts and trends hidden in databases. This explosion of data is relatively recent. In 2000, only a quarter of all the world's information was stored in digital format; the rest was stored on analog media such as paper. Today, however, more than 98% of all our information is digital[9].

Big Data (from now on BD) is a different phenomenon from the Internet itself, despite the fact that the latter has allowed its emergence since it has made it much easier, faster, and cheaper to collect and share data. If the Internet has completely revolutionized the way humanity communicates, the BD has represented a new way of processing information on a global level. The BD is usually a poorly defined concept: it can be understood simply as the accumulation of data large enough to have to be analyzed with large computers, but currently, we see how it can also be analyzed with desktop computers with standard software. Indeed, when we refer to the DB, we would have to take into account, not so much the extent of the data aggregation, but rather the ability to search, aggregate and cross different massive data sets[10], so that the potential of BD is multiplied through the use of cross technologies[11]. This explosion in data production and storage is relatively new. If we follow Montuschi, in 2000, only a quarter of the information stored globally was digital. The rest was preserved on paper and other analog media. But given that the amount of data in digital format expands so rapidly (it multiplies approximately by two every three years), this trend has been reversed: today, less than 2% of the information stored is not digital[12].

---

[7] Article 29 Working Party "Opinion 03/2013 on Purpose Limitation" (2013).

[8] Javier Puyol. “Big data and Public Administrations” [Conference]. International Seminar Big data for Official Information and Decision Making (2014).

[9] Kenneth Neil Cukier and Viktor Mayer-Schoenberger. “The Rise of Big data. How It's Changing the Way We Think About the World”. Foreign affairs Vol. 92, No. 3 (2013).

[10] Boyd, Danah and Crawford, Kate. “Critical questions for Big Data Information, Communication and Society” 662-679 (2012).

[11] Richard Cumbley and Peter Church. “Is “Big Data” Creepy? Computer Law & Security Review, 29, 601-609 (2013).

[12] Luisa Montuschi. “Troubled ethical issues in the information age, internet and the world wide web” CEMA Working Papers, University of CEMA (2005).

If we talk about data, information, or knowledge, we must bear in mind the differences between these concepts. The database often begins with the creation of unstructured data (raw data), which would not be an automatic synonym for information. Nor can the information be considered directly as knowledge. Previously it would have to be classified, or if you prefer structured (that is, processed in some way). It is from this processing that what we understand by knowledge would have to emerge. The data can exist raw, in a way that is not necessarily usable, and therefore would not have an autonomous meaning (by itself): the data must be located in a concrete context to become information, and if the context disappears, so will the information. In short, access to larger and larger amounts of information does not necessarily have to result in increased knowledge[13]. Therefore, in order to understand the BD, the role of the experts must also be highlighted, both in the technologies (which will determine the nature of the collection of this data), and in the procedures for its exploitation (the technicians in BD).

The use of large volumes of information required by the BD necessarily produces three profound changes in the way we perceive data. The first is to collect and use a large amount of data, rather than small amounts of samples, as has been done since statistical science for centuries. The second is that the benefits of using a lot of data of variable quality may be greater than those that provide us with smaller amounts of data that are more accurate. Third, in many cases, abandoning the investigation of the cause of phenomena in exchange for accepting correlations. Thus, we are collecting and analyzing massive amounts of information about events and everyone associated with them, looking for patterns that help predict them in the future[14].

This was anticipated by Alvin Toffler in "The Third Wave", referring to the birth of what this author will baptize as "the infosphere":

"When a problem arises we immediately try to discover its causes (...) when we approach a truly complicated problem (...) we tend to focus on two or three factors and to fear many others that, individually or collectively, can be far more important. (...) Because it can remember and interrelate a large number of causal forces, the computer can help us tackle such problems at a deeper level than usual. You can sift through vast masses of data to find subtle patterns, gather “glimpses,” and assemble it into larger, more meaningful units. (...) It can even

[13] Viktor Mayer-Schonberger and Kenneth Cukier. “The big data revolution” (2013)
[14] Idem.

suggest imaginative solutions to certain problems by identifying new, or hitherto unnoticed relationships between people and resources "[15].

In summary, we have to understand the concept of Big Data as the interaction between three realities/dimensions. First, a new technology, which tends to maximize computing power and at the same time perfect the algorithms to collect, analyze, link, and compare large amounts of data. Second, a new type of analysis derived from this technology, which consists of the identification in these data aggregates of patterns and trends of economic, social, technical, and legal demands. Finally, we are faced with new mythology: the widespread belief that BD offers a superior form of intelligence and knowledge that can generate previously impossible perceptions[16]. The exploitation of BD can have two aspects: one that deals with the population in general and one that focuses on the identification of behaviors that can be associated with specific individuals of this population.

In the first case, a clear example would be Google Flue Trends[17], the study carried out by Google, which facilitates inferring the real incidence of influenza disease globally from the analysis of searches carried out by users. Today, the information society services platform (PSSI) Google is one of the main providers of services and platforms for the dissemination of content on the Internet, most of them free[18]. In addition, in many cases the identification of the user is required to use the services, thus ensuring access to an enormous amount of cross information and at the same time multilevel. This has made Google today probably the maximum holder of information and personal data about Internet users, and therefore it is a key player in the BD phenomenon.

In the second aspect, the DB allows us to see (and infer) individual behavior patterns through various connections between variables contained in this data cloud[19]. An example of this second use is found in the case of companies that use these techniques to detect sectors of the population or markets more likely to consume their products, also including the design of personalized advertising (targeted advertising).

---

[15] Alvin Toffler. "The Third Wave" p179. New York: William Morrow and Company, Inc (1980).

[16] Boyd, Danah and Crawford, Kate. "Critical questions for Big Data Information, Communication and Society" 662-679 (2012).

[17] Fred O'Connor. "Google Flu Trends calls out sick, indefinitely". PCWorld. (20 August 2015).

[18] Maciej Szpunar. "Reconciling new technologies with existing EU law – Online platforms as information society service providers". Maastricht Journal of European and Comparative Law. 27. 399-405. (August, 2020)

[19] Luisa Montuschi. "Troubled ethical issues in the information age, internet and the world wide web" CEMA Working Papers, University of CEMA (2005).

In this way, we can see the BD from two perspectives. From a positive perspective, research with massive amounts of data can be a tool for analysis and knowledge creation, for better services and public goods[20]. It can also provide us with insight into political as well as community movements (both online and offline). From the negative perspective, one can suppose, from the simple invasion of personalized marketing, the violation of privacy rights, the reduction of civil liberties (such as protest or freedom of expression), as well as the increase of state control over individuals. Finally, several authors will see how, with the increasingly automated collection of information and its processing, it is necessary to know which are the systems that are guiding these practices, as well as the creation of a legal framework that adapts to the context each time more rapidly changing.

In conclusion, the concept of big data applies to all information that cannot be processed or analyzed using traditional tools or processes. The challenge is to capture, store, search, share, and add value to data that has been little used or inaccessible to date. The volume of data or its nature is not relevant. What matters is its potential value, which only new technologies specializing in big data can exploit. Ultimately, the objective of this technology is to provide and discover hidden knowledge from large volumes of data[21].

This phenomenon is based on the fact that there is currently more information around us than there has ever been in history, and it is being put to new uses. By way of illustration, we can point out that from the beginning of history until 2003, humans had created 5 exabytes (i.e. 5 billion gigabytes) of information. By 2021, we were creating roughly 2.5 quintillion bytes of data.[22].

In addition, big data makes it possible to transform into information many aspects of life that previously could not be quantified or studied, such as unstructured data (e.g., text data such as photographs, images, and audio files). This phenomenon has been dubbed as datafication[23] by the scientific community. Our location has been turned into data, first with the invention of longitude and latitude, and today with satellite-controlled GPS systems. Similarly, our words are now data analyzed by computers through data mining. And even our friendships and likes/dislikes are transformed into data, through social network relationship graphs or Facebook "likes".

---

[20] Idem.

[21] Javier Puyol. "An approach to big data". Journal of Law of the National University of Distance Education (UNED), No. 14 (2014).

[22] SeedScientific. website: https://seedscientific.com/how-much-data-is-created-every-day/#:~:text=How%20much%20content%20is%20created,2.5%20quintillion%20bytes%20of%20data.

[23] Kenneth Neil Cukier and Viktor Mayer-Schoenberger. “The Rise of Big data. How It’s Changing the Way We Think About the World”. Foreign Affairs Vol. 92, n.o 3 (2013).

Some people speak of the 1980s as the beginning of big data when processors and computational memory made it possible to analyze more information. However, the big data phenomenon is still humanity's latest step on an ancestral path: the desire to understand and quantify the world.

Today, new data are put to previously unknown uses, made possible by the growth of computer memory capacity, powerful processors, the incredible cheapness of collecting and storing this amount of information, and the development of mathematical analyses derived from traditional statistics. When we transform reality into data, we can transform information into new forms of value. One example of these new services is automated recommendation engines, which do not require an analyst to review the data and make these recommendations. For example, Amazon uses data it extracts from its customers to make recommendations based on previous purchases by other customers. And the professional network Linkedin also suggests people we might know by selecting a few data points from a massive amount of information about its more than 300 million members[24].

What are known as the three "Vs" of databases, variety, volume, and velocity were incompatible years ago, creating a tension that forced us to choose between them. In other words, we could analyze a large volume of data at high speed, but it had to be simple data, such as structured data in tables; in other words, the variety of the data had to be sacrificed. Similarly, large volumes of very varied data could be analyzed, but not at high velocity; the systems had to be left to work for hours, or even days[25].

However, with the emergence of big data these three attributes are no longer mutually exclusive but complementary:

Volume: big data involves collecting, storing, and processing large amounts of data and metadata[26], rather than studying a sample, as traditional statistics do. According to SeedScientific, we create roughly 2.5 quintillion bytes of data on a daily basis[27]. In fact, these amounts are so large that, for non-experts, they are meaningless, and adding more or fewer zeros to the figure does not even allow us to see the difference.

---

[24] Lutz Finger. "Recommendation Engines: The Reason Why We Love Big data". Forbes Tech (September 2, 2014).

[25] Paulo Goes. "Big Data and IS Research". MIS Quarterly, 38(3), iii-viii. (2014).

[26] Metadata is data that describes other main data, with which they are associated. For example, a digital photograph taken with a mobile device may contain other information in the form of metadata such as the location or the date the image was taken.

[27] SeedScientific. website: https://seedscientific.com/how-much-data-is-created-every-day/#:~:text=How%20much%20content%20is%20created,2.5%20quintillion%20bytes%20of%20data.

This volume of data is so large that it can no longer be analyzed using traditional tools and processes such as MS Excel or SQL. It has been necessary to start using new systems, such as NoSQL or Apache Hadoop software, which allows millions of bytes of information to be processed and organized into thousands of nodes.

Velocity: the velocity at which data is created and processed is constantly increasing, and it is often important for organizations to be able to analyze it very quickly, even in real-time, something that is sometimes impossible with traditional systems. Big data makes it possible to transfer data cheaply and efficiently so that both dynamic data that is being created and static or historical data that has already been stored can be analyzed.

For example, real-time data analysis can help track the path of hurricanes and their intensity. This could make it possible to make predictions about where damage may occur hours or even days in advance.

Variety: the data collected comes from both structured and unstructured sources: banking transactions, satellite images, social networks, website content, mobile geolocation devices and thousands of applications, the connections of the internet of things, web 2.0 services, and even the human body (e.g. when using biometric identification systems). Currently, roughly 10% of our data comes from structured sources, while the remaining 90% is unstructured data[28].

Extracting information from such diverse data is a major challenge. The technologies that have been developed for big data make it possible, among other solutions, to combine data even though they are not stored in files with the same structure. For example, a retail chain can analyze sales data together with temperature data to create a real-time predictive model for each of its stores.

Some experts consider that of all the "Vs", variety is the most relevant feature of big data[29]. This is because, for example, if a company wants to extract information from its own customer database, even if it is very large, it may not need to use new analysis tools or face new privacy issues. However, when the company wants to combine that data with other external sources, then it will be carrying out completely different activities that could be called big data.

[28] Bernard Marr. "What Is Unstructured Data And Why Is It So Important To Businesses? An Easy Explanation For Anyone" [Website]. Bernard Marr & Co (2013) https://bernardmarr.com/what-is-unstructured-data-and-why-is-it-so-important-to-businesses-an-easy-explanation-for-anyone/

[29] Information Commissioner's Office (ICO, UK data protection authority). Big Data and Data Protection (2014).

These "three Vs" can be extended with three more: veracity, visualization, and value of the data[30].

Veracity: veracity refers to the level of reliability or quality of the data. Achieving high-quality data has become a challenge, especially important when dealing with unstructured data. However, as IBM asserts[31], some data is uncertain by nature, such as sentiment, the future, GPS sensors bouncing between skyscrapers in a city, or data created in human environments such as social networks; and no amount of data cleansing can correct it. Thus, managing uncertainty is an essential issue when dealing with big data technologies.

And as IBM proposes, to manage uncertainty, analysts need to create context around the data. One possible way to do this is to combine various data sources to result in more reliable information. For example, when social media comments are combined with geolocation data to analyze people's reactions and the impact of an event such as a large concert, election rally, or soccer match.

Visualization: being able to visualize data is essential to understand it and make decisions accordingly.

For example, the use of big data techniques makes it possible to combine data and obtain a prediction of when and where certain types of crimes will occur (after a large soccer match, etc.). However, an endless list of pages with coordinates showing where crimes occur is not manageable. Visualization tools using, for example, maps in which the color intensity shows the probability of each type of crime occurring can be crucial to really understand the data.

Value: the ultimate goal of big data processes is to create value, whether understood as economic opportunities or innovation. Without it, efforts become meaningless.

Apart from these characteristics, big data can also be characterized in terms of its differences from traditional processing tools, such as the intensive use of algorithms or the use of data for new purposes. Before big data, to analyze a data file it was necessary to first decide what one wanted to study and what one expected to find, i.e., to establish a hypothesis, in order to launch a search and identify the relevant data associated with those parameters. Sometimes, when the analysis was more complex, numerous algorithms could be run on the data to find correlations. So what does big data bring to the table again?

---

[30] Paulo Goes. “Big Data and IS Research”. MIS Quarterly, 38(3), iii-viii. (2014).

[31] IBM Institute for Business Value, in collaboration with the Said School of Business at the University of Oxford. "Analytics: Using Big Data in the Real World". IBM Global Business Services (2012).

Finding a correlation means finding a phenomenon that we did not know about, and this means finding new information. This new information can be fed back into the algorithms to make them perform more accurate searches so that the previous results refine the operation of the algorithm and the system "learns". This is what is known as "machine learning" (which we could translate as computational learning)[32]. Well, the use of algorithms in this way is a novelty of big data.

The data analysis system combines computational learning techniques and natural language processing, so that data from a particular patient can be entered and, for example, a prediction can be obtained that determines the likelihood of a particular treatment being successful.

The use of data for new purposes: big data analytics often reuses data that was obtained for a first purpose, and gives it a new purpose. This is a consequence of all of the above. If analyzing more and more data allows us to learn more information, organizations or governments can identify previously unknown problems that can be understood or addressed with this information.

For example, imagine a wind power plant that installs smart sensors on the windmills to monitor their operation. The company's goal is to be able to manage more efficiently how many technicians will be needed in each area to check for faults. However, the data on mill faults can also be used for other purposes. For example, the company can observe which parts fail most frequently, and this data can in turn be combined with data from suppliers to consider purchasing those failed parts from another manufacturer. Or it can be learned which failures occur more often in hot, dry weather versus cold, wet weather, and the company can manage its inventories more accurately.

All these changes are bringing about a paradigm shift in the way information is analyzed. This paradigm shift is embodied in three major trends that have been illustrated by Kenneth Neil Cukier and Viktor Mayer-Schöenberger[33].

First is the shift from "something" to "everything". Traditionally, the way in which data were treated was by means of representative samples of reality. We relied on small amounts of information that could be easily handled to explain complex realities. In general matters, samples and statistics work well, but when we want to draw conclusions from specific

[32] Batta, Mahesh. (2019). Machine Learning Algorithms -A Review. (January, 2019)

[33] Kenneth Neil Cukier and Viktor Mayer-Schoenberger. “The Rise of Big data. How It’s Changing the Way We Think About the World”. Foreign Affairs Vol. 92, n.o 3 (2013).

subgroups of the sample, statistics are no longer reliable[34]. This is because random samples are sufficient to describe global realities, but not to detect particular behaviors of subgroups. For example, statistics are able to give an answer to the question of which electoral candidate is preferred by single women under 30 years of age. Statistics forces us to know a priori what we want to analyze and to choose a sample accordingly. But if once the survey has been carried out, we want to re-analyze a subgroup of this population, for example, the preferred candidate of single women over 30 years of age, with a university education, Spanish nationality, and foreign parents, the conclusions will not be valid. This is because in the chosen sample there is probably not a sufficiently large group with these characteristics to be able to draw conclusions. However, if we collect information on a massive scale, this problem disappears. We do not need to know beforehand what we want the information for, we simply collect as much information as possible and in this way, we can analyze the behavior of the main group as well as of the various subgroups we want to create a posteriori.

Secondly, the change from "clean" to "chaotic". If we tend to collect such a large amount of data, we have to give up trying to make all this information structured and clean, and we have to accept some disorder. The benefits of analyzing large quantities outweigh the disadvantages of allowing these small inaccuracies (provided the data are not completely incorrect). Viktor Mayer Schöenberger gives machine translators as an example[35]. Their beginnings date back to the 1990s when the Canadian government needed an efficient means of translating its writings into English and French. At that time, IBM devised a statistical translation system that inferred which word in the other language was the best alternative for translation. Today, Google has taken over. Its translator now supports more than 90 languages, ranging from the most widely used languages to others such as Sinhalese and Kazakh. Google's translation engine is not based on a few perfect translations, but on huge amounts of data from a wide variety of sources: corporate websites of companies, European Union documents in all their translated versions, and so on. The results of its translations are not perfect, but they are certainly useful in a large number of languages.

Thirdly, there is a shift from "causation" to "correlation". It is no longer so much important to discover the causality between two facts, but their correlation[36]. Thus, instead of trying to understand exactly why a machine breaks down or why the side effects of a drug disappear, big data allows researchers to collect and analyze massive amounts of data on these

[34] Idem.
[35] Idem.
[36] Idem.

events and everything associated with them, to find relationships between variables to uncover hidden patterns and predict when these events might happen again. In any case, it should be borne in mind that this approach to reality also entails risks, which will be analyzed later in this paper.

1.2 Potential benefits of big data

In this context, it is clear that the opportunities generated by big data are enormous, and these opportunities are already today, in many cases, a tangible benefit.

The digital universe is a business area that is absolutely on the rise and will have enormous value in the future. Some of the most relevant benefits of big data are to be able to offer an increasingly accurate view of the fluctuations and yields of all types of resources, to enable experimental adaptations to be made at any scale of a process, and to know its impact in almost real-time, to help to better understand demand and thus make a much tighter segmentation of the supply for each good or service, or to accelerate innovation and the provision of increasingly innovative and more efficient services[37].

Large companies were able to see the potential value of big data and data mining techniques years ago, and so Axciom, Google, IBM, and Facebook have been investing for years in discovering new uses for data, how to process it, and how to transform it into value. Following the great pioneers, in most sectors, both mature companies and new entrants are implementing strategies to innovate and capture value. For example, in the healthcare sector, some pioneering companies are analyzing the health outcomes of certain widely prescribed drugs, and are discovering benefits and risks that were not uncovered during clinical trials. Other companies are collecting data from sensors embedded in products such as children's toys or industrial goods to find out how these products are being used in practice. With this new knowledge, companies are able to generate new services and design future products. In this way, data analysis becomes an important competitive advantage for companies[38].

A concrete example in the retail sector is the Walmart supermarket chain[39], which collects data on its customers' purchases that it then analyzes to understand their consumption habits. With the millions of bytes of information it holds, the company decided to try to make

[37] Javier Puyol. “Big data and Public Administrations” [Conference]. International Seminar Big data for Official Information and Decision Making (2014).
[38] Wang, Lidong and Alexander, Cheryl. Big Data Analytics in Healthcare Systems. International Journal of Mathematical, Engineering and Management Sciences. 4. 17-26. (January, 2019)
[39] “How Big Data Analysis helped increase Walmarts Sales turnover?”[Website]. (25 January, 2022) https://www.projectpro.io/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109

sales predictions in certain circumstances, such as in hurricane warning situations. The analysis of the data uncovered such surprising patterns as the fact that the star product that consumers buy before hurricanes is beer, or that when a hurricane threatens, sales of strawberry "Pop-Tarts" soar up to seven times higher than ordinary sales. With this knowledge, the chain is supplied before a hurricane, and this information is not only power but also money.

And it is not only companies that use mass data to make a profit. For public administrations, the use of big data can lead to faster and more effective decision-making, predictive analysis, and continuous improvement of work systems, as well as improving efficiency in sensitive issues such as citizen protection or healthcare.

The New York City Council has used big data analytics for purposes as diverse as preventing traffic jams since traffic regulation in large cities is a cause of related problems such as the difficulty of attending to victims of fires or the inefficiency of city services[40].

In Spain, many companies have also embraced the benefits of big data. Telefónica is devoting significant efforts to developing lines of research using big data techniques. For example, in May 2014 Telefónica R&D published a report prepared jointly with RocaSalvatella on tourism in the cities of Madrid and Barcelona, based on large amounts of data from two different companies, Telefónica Móviles Spain and BBVA[41]. Specifically, the report cross-referenced data from foreign terminals that used Telefónica's infrastructure with data from electronic payments by foreign cards that used BBVA's infrastructure. In order to publish the conclusions of the report while guaranteeing user privacy, the data were previously anonymized, aggregated, and extrapolated using statistical techniques.

Similarly, the company O2, a subsidiary of Telefónica in the UK, also claims to cross-reference large amounts of data from sources such as payment histories, social networks, which companies customers call, consumer preferences and segmentation, etc. In an internal session led by Dave Watkins, Director of Strategic Analytics and Business Intelligence at O2 in the UK, Telefónica's big data opportunities are highlighted. Analytics would, they say, allow them to add value to customers by studying how Telefónica products are used, obtaining the location of users, their television viewing habits, etc[42]. This would enable them to derive valuable insights into how Telefónica's products are used. This could infer valuable knowledge for the company's strategy, such as mobility patterns or customers' social circles. Population

[40] Isabella Fish. “Drivers' phone data is being used by York City Council to try and ease the city's traffic jams and could be extended across the country”. Daily Mail. (19 April, 2018)
[41] Hernandez, Emilcy & Duque, Néstor and Cadavid, Julián. “Big Data: an exploration of research, technologies and application cases”. 20. 17-24. (December, 2017)
[42] Idem.

movement matrices can also be generated or population density maps, estimated by call concentration. This could, in turn, help extract valuable data for other social actors such as government agencies. O2 estimates that the incremental benefits it could gain from using big data would amount to £434 million.

However, as Telefónica points out, the Achilles heel of data-driven companies is privacy, and the reputational risk they face is very high. Large companies such as Google, Facebook, AOL, and Microsoft are among the worst perceived by users in terms of privacy. As we will see below, the defense of privacy and data protection is one of the most important challenges facing big data today.

In short, big data helps to create new business opportunities and even new markets and new categories of companies. It is normal that many of these new companies are in the middle of the data flows, capturing and analyzing information about products and services, suppliers and customers, or consumer preferences.

What's more, some of the most important opportunities for creating value from personal data are still unknown. Much of the information captured currently resides in silos separated by different legislative rules and contracts, and the lack of an effective system for data transfer prevents the secure creation of value. However, it is necessary to move data to produce value. "Data by itself sitting on a server is like money under the mattress. It is safe, but stagnant and underutilized"[43].

In addition to the enormous opportunities presented by big data, it should not be forgotten that it also has certain limitations, which we will see below.

[43] World Economic Forum and the Boston Consulting Group "Rethinking Personal Data: Strengthening Trust" (2012).

1.3 Risks of big data

Indeed, big data must face certain challenges or constraints. Specifically, some of the most important challenges (leaving aside the technical difficulties of storage or computational research) are (i) the risk of falling into erroneous conclusions; and (ii) the risk to individuals of making automated decisions without human bias; and (iii) the risk to the privacy of individuals[44]. In this section, we will analyze the first two risks briefly, and then focus the rest of the chapters of the research on the problems that big data poses for privacy and data protection.

(i) The risk of falling into erroneous conclusions

One of the fundamental ideas of big data is that the massive analysis of past data can locate patterns that allow future predictions to be made. But after analyzing the data, it is important to find the true relationship between variables in order to create a predictive model. In other words, it is essential to be able to differentiate causality from chance. Sometimes we tend to confuse the two concepts, but although it may seem like a tongue twister, differentiating them has very important practical consequences. In fact, causality is the area of statistics most misunderstood and misused by non-specialists[45]. So before we can establish this difference, we will begin by reviewing the concept of statistical correlation in a very simple way.

In the statistical sciences, correlation is the degree of relationship between two variables. That is, two variables are said to be correlated when an increase or decrease in one causes a clear change in the other. Thus, if the increase in one value is accompanied by an increase in another value, there will be a positive correlation. If the increase in one value causes a decrease in another value to be observed, we have a negative correlation. And if, despite a change in one value, we observe no change in another value, there is a zero correlation. For example, there is a positive correlation between the number of hours a person studies and the grade he or she obtains on an exam. A relationship has also been found between a country's GDP level and the average penis size of its inhabitants.

Well, when two variables are correlated, it is possible that they also have a causal relationship. This implies that one event is a direct consequence of the other, or, in other words,

[44] Bottles, K., Begoli, E., & Worley, B. (2014). Understanding the Pros and Cons of Big Data Analytics. Physician Executive, 40(4), 6-12.
[45] Clarke, Roger. “Big data, big risks. Information Systems Journal”. 26. 77-90. (January, 2016)

that there is a cause-effect relationship in such a way that the occurrence of the first event (which we call cause) causes the second (which we call effect). In the above example, there is a causal relationship between the number of hours of study and the result of an exam. However, a correlation between two variables does not always imply causality.

In fact, sometimes two variables are correlated, even if this correlation occurs by chance. This type of relationship is called spurious or false[46]. For example, it is a mere chance that the GDP level of a country is correlated with the penis size of men in that country.

Thus, statistical data may show a correlation, and this is a good starting point; but after that, it is up to us to add a subjective approach and study whether there is indeed a pattern between the two variables that explains a true connection, or whether it is just a coincidence.

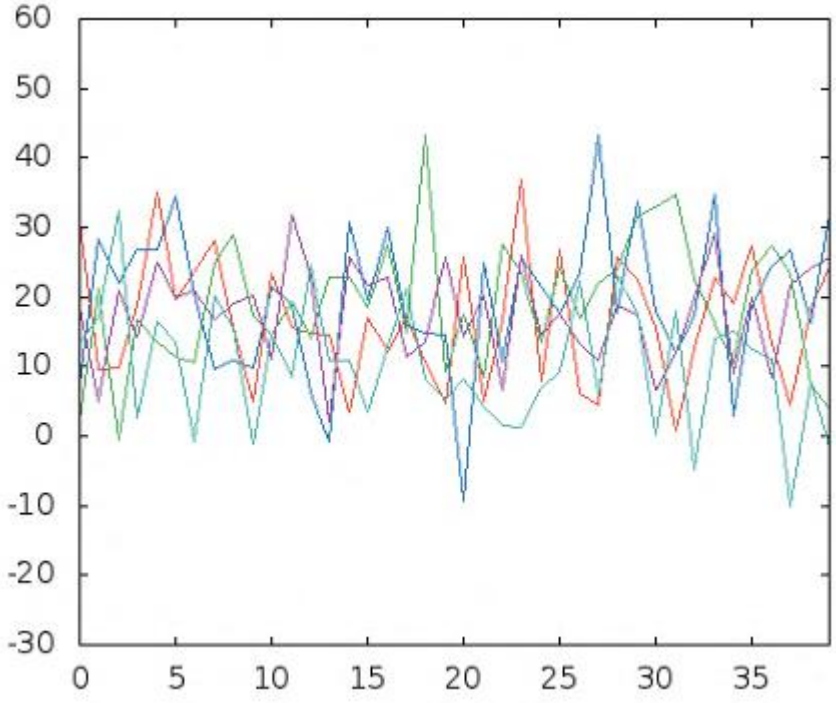
Let us take another example: years ago it was observed that there was a correlation between cancer cases and tobacco consumption. At first, it was not known whether smoking really caused an increase in the probability of suffering cancer, so medical research had to be initiated to determine that there was indeed a cause-effect relationship between the two variables.

Specifically, we can find two types of error in the interpretation of spurious relationships, error by chance and error by fusion[47]. But if this has always been a problem that needed to be taken into account, why is it especially important when talking about big data?

First, let's look at random errors. The statistician Stanley Young has long been warning of what he has called "the tragedy of large data sets": the more variables you study in a large data set, the more correlations that may be evidence of pure or spurious statistical significance. Thus, the more data we have, the more likely we are to find illusory relationships without any real significance, even if both have a strong statistical relationship. Similarly, using the Monte Carlo method to generate random variables, it can be seen that spurious correlations grow exponentially with respect to the number of variables. This means that, if misinterpreted, the analyst may end up being misled by the data.

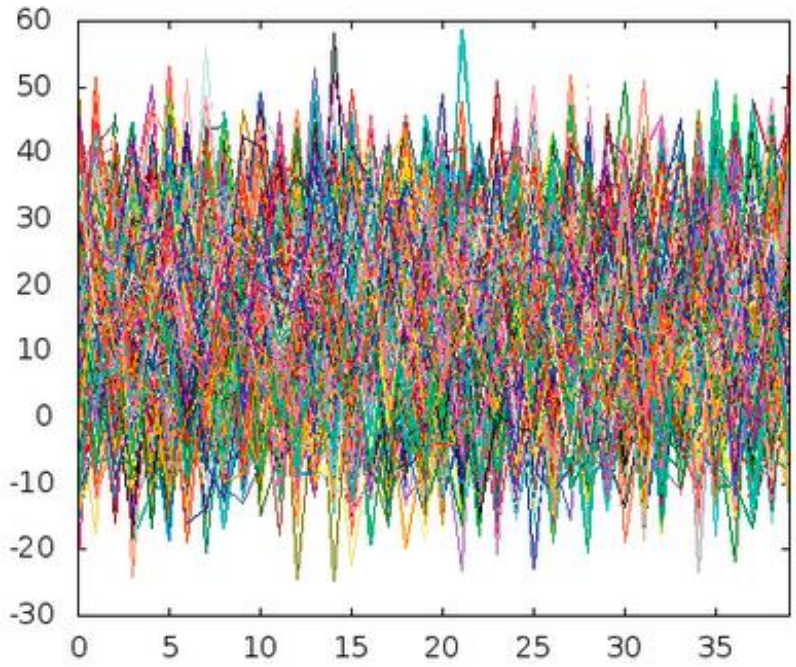
Ricardo Galli, a computer scientist, and free software activist conducted the following experiment[48]. Suppose we have the following data on the evolution of five economic variables over the last few years.

[46] Idem.
[47] Idem.
[48] Ricardo Galli. "Be careful with Big Data." Free Software Blog internet, legal (May 29, 2013).



Graph: Variables 1 to 5[49]

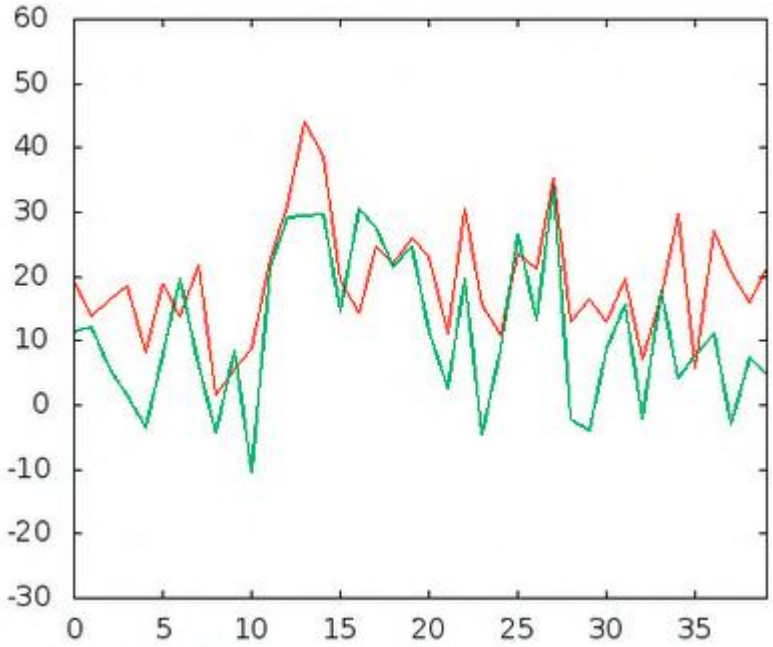
This small amount of data does not show us any correlation between variables. But now suppose that, instead of five variables, we can analyze a thousand variables (something similar to what would happen with big data). Our graph would look something like this:



[49] Idem.

Graph: Variables 1 to 1000[50]

Analyzing so much data, we find variables that are positively correlated (i.e., they grow and decline in tandem), for example, those presented in the following graph. In other words, when we have increased the number of variables in the study, the number of correlations that we can observe between these variables has increased.



Graph: Two variables show a positive correlation[51]

The analysis of massive data has made it possible to detect correlations that we did not know about, which in turn may allow us to discover other information that we do not know about. The problem is that this relationship is due to chance.

Indeed, these data were artificially created by Galli. In his experiment, he explains that he generated these economic variables with pseudo-random and independent numbers so that the conclusions obtained are also random.

Imagine the consequences that a misinterpretation of economic variables can have on a country's decision-making, or on econometric sciences.

As we have already mentioned, what happens in these series is what is known in statistics as random error; the relationship between the two variables is pure coincidence.
Thus, at a time when we are storing and analyzing massive amounts of data, we are exposed to finding more spurious relationships than ever, which, if not questioned, will lead us to erroneous conclusions. Because of today's seemingly blind trust in data, such erroneous

[50] Idem.
[51] Idem.

conclusions can determine decisions about people, strategic business actions, or government policies that are based on pure chance.

Finding the causes of a given event is what allows us to know how reality works and to predict how it will work in the future. However, the way to find the real cause of a relationship is through controlled experiments. The problem is that, in most cases, these are too expensive, time-consuming, or even technically impossible to carry out. This is why this may be an important area of research in the coming years[52].

(ii) The risk of automated decision-making

The evolution of data science makes us question when the intervention of a person is needed to supervise the conclusions obtained in an automated way before they are transformed into decisions; decisions that can cover spectrums as broad as advertising, the granting of a loan or a medical diagnosis.

Many of the operations carried out on the Internet are based on automated decision-making without human intervention, except, obviously, prior intervention to set the parameters for the adoption of the automated decision[53]. In other words, there is human intervention when the algorithms that will analyze the data to make a decision are created, but on many occasions, there is no human control to check the decision.

Basic research and intuition are being usurped by algorithmic formulas[54]. Steve Lohr states that, in fact, automated decision making is designed to take humans out of the equation, "but the impulse to want a person to monitor the results that the computer spews out is very human"[55]. Because our logic is different from that of machines, we need to feel that correlations translate into causes. We have always functioned this way: motivating our conclusions is one of the issues to which we devote most time in our analyses. This scheme falls apart if big data tells us what to do without further justification.

---

[52] Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, Bernhard Scholkopf. "Distinguishing cause from effect using observational data: methods and benchmarks", versioon 2. Journal of Machine Learning Research, Cornell University (2015).

[53] Ana Victoria Sanchez Urrutia, Héctor Claudio Silveira Gorski, Mónica Navarro Michel, Stefano Rodota. "Technology, privacy and democratic society". *Icaria* (2003).

[54] Anderson, Richard. "How Mathematicians Dominate the Markets". BBC Economy (October 1, 2011).

[55] Steve Lohr. "If Algorithms Know All, How Much Should Humans Help?" The New York Times News Services (6 abril 2015).

Blindly trusting algorithms often leads to companies making decisions about us without us being able to know why they have made them. This is one of the workhorses of big data.

Many consider the marketing sector to be the ideal place to test and correct new mathematical and technological tools, as it has few risks and many benefits. In marketing, a mistake simply means that a consumer sees the wrong ad, and a hit can lead to increased sales.

However, concern increases when these new techniques begin to be used in other sectors, such as banking, insurance, and, above all, healthcare. In these, serious doubts arise as to when human intervention is really necessary to overcome the results achieved by algorithms. One tendency is to keep humans involved in the decision-making process, but others believe that this would be counterproductive[56].

In the banking sector, for example, many companies are turning to big data analysis, following the basic banking principle of "know your customer". Data analysis makes it possible to know borrowers better than ever before and to predict whether they will repay their loans more accurately than if only their credit history is studied. This system relies on algorithms that analyze data in a complex and automated way (and even its advocates have doubts about the process)[57].

Aware of this, IBM has created a supercomputer, called Watson, which is being tested in the healthcare sector. It is capable of reading thousands of documents per second, a speed unmatched by humans, looking for correlations and other important insights. The Watson Paths program, in particular, allows physicians to see the evidence and the path of deductions that the computer has followed to draw its conclusions (e.g. how it has concluded the medical diagnosis it makes). This pioneering experiment attempts to provide a solution to this lack of human supervision of the results of the algorithms, to the lack of a machine-human translation, which will undoubtedly continue to advance as the science of big data progresses. The Watson computer is not the only initiative in this direction. Some companies claim that their employees review the recommendations made by their computers, "although it is rare for them to reject what the algorithms dictate"[58].

To date, the reality for organizations that rely on data science to make decisions is that they are rarely reviewed. In addition, there are also those in favor of not giving the human being veto power over decisions made analytically through algorithms. They claim that this would

[56] Idem.
[57] Idem.
[58] Idem.

introduce a human bias into a system in which one of its virtues is precisely that it promises decisions based on data and not on intuition or arbitrariness. Thus, the results provided will be better.

Faced with this dichotomy, one possible solution already suggested by analysts is to program or tweak the algorithms in such a way as to provide greater protection for individuals, and thus reduce the risk of making the wrong decision about a specific person. Thus, "the aim is not necessarily for a human to foresee the result a posteriori, but to improve the quality of the classification of individuals a priori "[59].

For example, continuing with our mention of the banking sector, this could translate into the algorithms used by companies being adjusted so that the probability of an individual being associated with a doubtful payment profile is lower.

Having reviewed the risks that big data imposes in relation to making erroneous decisions based on spurious relationships, and the risk of automated decision making without human supervision, it remains to analyze the risk that big data poses to the privacy and data protection of individuals. The remainder of this paper will focus on this problem from a legal point of view.

[59] Idem

Chapter 2: Big Data and Data Protection

2.1 The impact of big data on data protection regulations

Big data may represent a challenge for different bodies of law, such as data protection, prohibition of discrimination, civil liability, competition law, intellectual property rights, etc. This study focuses on privacy and data protection issues[60].

It is important to point out that data protection law applies when the information on natural persons makes them identifiable or identifiable. However, when the data does not make a person identifiable, this regulation does not apply. In other words, when data are rendered anonymous through anonymization techniques, they become non-personal data, and the privacy of individuals is protected so that no data protection regulation needs apply. Along with anonymization, our standard also deals with what it calls the dissociation process, which allows the creation of pseudonymous data, a category of data that, without being anonymous, has more privacy safeguards than purely personal data.

Big data challenges data protection rules by facilitating the re-identification of subjects, not only on the basis of pseudonymous data but also on the basis of data that we considered anonymous. In other words, anonymization techniques are no longer always sufficient with the advent of big data. This means returning to the basic debate as to which data are personal and which are non-personal. All these concepts will be discussed in detail below.

To sum up, big data threatens data protection regulations, due to several reasons[61]:

i. The principle of "data minimization"[62] is not met in practice. This principle implies that the data collected should not be excessive, but that only the minimum amount necessary for the purpose for which it is collected should be collected. However, in very few cases do data protection authorities effectively oblige companies to redesign their processes to minimize the data collected. Moreover, the principle of data minimization runs againts the very logic of

[60] Kuner, Christopher & Cate, Fred & Millard, Christopher & Svantesson, Dan. (2012). The challenge of 'big data' for data protection. International Data Privacy Law. 2. 47-49.
[61] Idem.
[62] The data minimisation principle is expressed in Article 5(1)(c) of the GDPR, which provide that personal data must be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed".

big data. The new analytical models are based precisely on the study of massive amounts of data without which the knowledge that big data allows us to extract could not be extracted.

ii. The regulations rely too much on the individual's informed consent to collect and process his or her personal data[63]. This is a problem, given the experience that the vast majority of individuals do not read privacy policies before giving their consent; and those who do so do not understand them. Thus, granting consent is, in general, an empty exercise.

iii. Anonymization has proven to have limitations. Although it was presented as the best solution for processing data while protecting the privacy of the subjects, over the years there have been numerous cases of re-identification of databases that had been anonymized. It is becoming easier and easier to re-identify subjects, not only through the analysis of different sources containing partial personal data of a person but also through non-personal data[64]. This implies a weakening of anonymization as a measure to ensure privacy during data processing.

iv. Big data increases the risk associated with automated decision-making. This means that life-altering decisions, such as calculating our credit risk, are subject to automatically executed algorithms. The problem arises when the data that are analyzed by algorithms are not accurate or truthful, but individuals have no incentive to correct them because they are not aware that the data are being used to make decisions that affect them.

[63] Ira S. Rubinstein. “Big data: The End Of Privacy Or A New Beginning?”. International Privacy Law, Vol. 3, n.o 2 (2013).
[64] Fred H. Cate. “The failure of Fair Information Practice Principles”. Consumer Protection In The Age Of Information Economy (2006).

2.2 The legal framework and characteristics of consent

Express consent is consent that is given explicitly, concretely, and directly and is recorded in one or more ways so as to leave no room for doubt. One of the most frequent ways of granting it is by signing an official document so that the interested parties can consult it if there is any disagreement in the future.

We can find a definition of express consent in Article 4.11 of the GDPR, which includes this principle as: “any freely given, specific, informed and unambiguous indication of the data subject’s wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her”.

As mentioned above, since the entry into force of the GDPR, only the express consent of the data subject is valid for any processing of his personal data, including receiving commercial communications or transferring his data to third parties.

According to the GDPR, express consent "requires a clear and unequivocal statement by the data subject that he/she allows or consents to the processing or transfer of which he/she is informed, through the declaration of his/her will, which may be made in writing, verbally, through telematic notification or by any other means"[65].

This consent exists when the user must explicitly say yes (example: "if you want me to send you commercial information, check this box") or no, but in any case, must perform an action. Therefore, in the GDPR tacit consent is no longer sufficient.

According to the regulation express consent must have the following characteristics in order to be considered legitimate[66]:

- The controller must be able to demonstrate that the data subject consented to the processing of his or her personal data.
- If the consent is given in a written statement that also refers to other matters, the request for consent must be clearly distinguishable from the other matters, intelligible, and easy to understand.
- Consent may be withdrawn by the data subject at any time.
- The data subject must give his or her consent freely, i.e. the provision of a service or the delivery of a good may not be made conditional on the granting of consent for the processing of personal data that are not necessary for that purpose.

[65] See Article 7 GDPR. “Conditions for consent”.
[66] Idem.

Consent can only be valid if the data subject can make a real choice and there is no risk of deception, intimation, or significant negative consequences if he or she does not consent.

Sometimes consent is not freely given. This is often because there is a subordinate relationship (such as an employer-employee relationship) between the data subject and the data collector and processor. On other occasions, the lack of freedom may be due to some kind of social, financial, or psychological coercion[67].

Moreover, to be valid, consent must be specific. Thus, GDPR states[68] that indiscriminate consent without specifying the exact purpose of the processing should not be admissible.

To be specific, consent must be comprehensible; that is, it must refer clearly and precisely to the scope and consequences of the data processing. It cannot refer to an indefinite set of processing activities. This implies knowing a priori what data are collected and the reasons for the processing.

Example: social networks

Access to social networking services is often subject to authorization for different types of processing of personal data[69].

The user may be asked to consent to receive behavioral advertising before being able to sign up for the services of a social network, without any further specification or alternative options. Considering the importance that social networks have acquired, certain categories of users (such as teenagers) will agree to receive such advertising to avoid the risk of being excluded from social interactions.

In this context, GDPR is of the opinion that the user should be able to give free and specific consent to receive personalized advertising, regardless of his or her access to the social network service.

However, in practice, the user is often prevented from using an application if he does not consent to the transmission of his data to the application developer for various reasons, including behavioral advertising and the resale of data to third parties. Since the app can

[67] Konow James. "Coercion and Consent". Journal of Institutional and Theoretical Economics JITE. (Marhc, 2014)

[68] See Article 7 GDPR. "Conditions for consent".

[69] Nunan Dan and Yenicioglu Baskin. "Informed, Uninformed and Participative Consent in Social Media Research". International Journal of Market Research. 55. 791. (January, 2014)

function without the need to transmit any data to the app developer, GDPR advocates differentiated user consent for different purposes. Different mechanisms could be used, such as drop-down sales, to give the user the possibility of selecting the purpose for which consent is given (transmission to the developer, value-added services, personalized advertising, transmission to third parties, etc.).

On the other hand, there are others who consider that sending behavioral advertising is a power of the data controller, which should be detailed in the terms and conditions of use of the platform. Each individual would therefore have the option of accessing the service and receiving advertising, or not accessing the service unless it is an essential service.

The specific nature of consent also means that if the purposes for which the data are processed by the controller change at any time, the user must be informed and be in a position to consent to the new data processing. The information provided should mention the consequences of refusing the proposed changes.

Particularly relevant in big data is the fact that consent must apply to a given context, as well as the fact that if the purpose for which the data will be used changes, consent may need to be sought again.

And this is precisely because the value of big data lies in the fact that the new information that is created allows new uses to be made of the data. It is precisely in these secondary uses that the potential of big data lies. This way of conceiving consent would mean that every time a new use for the data is discovered, the data controller would have to ask for the consent of each of the individuals whose data is being processed a second time. This may in many cases be technically unfeasible, not to say that companies would not be able to bear the costs.

GDPR enforces that in order to be valid, consent must be informed. This implies that all necessary information must be provided at the time consent is sought, in a clear and comprehensible manner, and must cover all relevant issues and transparency is a must. This is sought after according to Article 7 of GDPR.

Consent as an informed expression of will is particularly important in the context of transfers of personal data to third countries, insofar as it requires the data subject to be informed about the risk of his or her data being transferred to a country lacking adequate protection.

According to EDPB (European Data Protection Board) for the consent to be considered informed several requirements have to be met:

I. the controller's identity,[70]

II. the purpose of each of the processing operations for which consent is sought,[71]

III. what (type of) data will be collected and used,[72]

IV. the existence of the right to withdraw consent,[73]

V. information about the use of the data for automated decision-making in accordance with Article 22 (2)(c)[74] where relevant,

VI. on the possible risks of data transfers due to the absence of an adequacy decision and of appropriate safeguards as described in Article 46.[75]

As mentioned above, according to the GDPR, consent must also be unambiguous. In other words, the procedure by which consent is given must leave no room for doubt as to the data subject's intention to give consent. If there is any doubt as to the subject's intention, an equivocal situation will arise.

This requirement obliges data controllers to create rigorous procedures for individuals to give their consent. This means either seeking express consent or relying on procedures that allow individuals to express clear inferable consent. The Art. 29 WP and the European Data Protection Supervisor (EDPS), have stated in their contributions to the discussions on the new data protection framework that:

"It is not always easy to determine what constitutes true and unambiguous consent. Certain data controllers exploit this uncertainty by resorting to methods that exclude any possibility of giving true and unambiguous consent"[76]

Example: online gambling[77]

---

[70] See also Recital 42 GDPR: " […]For consent to be informed, the data subject should be aware at least of the identity of the controller and the purposes of the processing for which the personal data are intended.[…]."

[71] Again, see Recital 42 GDPR.

[72] See also WP29 Opinion 15/2011 on the definition of consent (WP 187) pp.19-20.

[73] See Article 7(3) GDPR.

[74] See also WP29 Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (WP251), paragraph IV.B, p. 20 onwards.

[75] Pursuant to Article 49 (1)(a), specific information is required about the absence of safeguards described in Article 46, when explicit consent is sought. See also WP29 Opinion 15/2011 on the definition of consent (WP 187)p. 19.

[76] Opinion of the European Data Protection Supervisor of 14 January 2011 on the Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - "A comprehensive approach to the protection of personal data in the European Union."

[77] Gainsbury, Sally & Angus, Douglas & Procter, Lindsey & Blaszczynski, Alex. "Use of Consumer Protection Tools on Internet Gambling Sites: Customer Perceptions, Motivators, and Barriers to Use". Journal of Gambling Studies. (March, 2020)

Imagine a situation in which the provider of an online game requires players to provide their age, name, and address before participating in the game (in order to make a distribution of players by age and address). The website contains an advertisement, accessible through a link (although access to the advertisement is not for participating in the game), which indicates that by using the website and thus providing information, players consent to their data being processed for the online game provider and other third parties to send them commercial information.

In the view of Art. 29 WP, accessing and participating in the game does not amount to giving unambiguous consent to the further processing of personal information for purposes other than participating in the game. Such behavior does not constitute an unambiguous manifestation of the individual's wish to have his or her data used for commercial purposes.

Example: default privacy settings[78]

The default settings of a social network, which users do not necessarily access when using it, allow, for example, the entire "friends of friends" category to see all the user's personal information, and users who do not want their information to be seen by friends of friends have to click a button. The file manager considers that if they refrain from acting or do not press the button they have consented to their data being visible. However, it is highly questionable whether not pressing the button means that people generally consent that their information can be seen by all friends of friends.

Because of the uncertainty as to whether inaction means consent, the Art. 29 WP considers that failure to click cannot be considered unambiguous consent.

[78] Baek, Young Min & Bae, Young & Jeong, Irkwon & Kim, Eunmee & Rhee, June. “Changing the default setting for information privacy protection: What and whose personal information can be better protected?”. The Social Science Journal. 51. (July, 2014)

Article 4(11) of the GDPR builds upon this definition built by the Art. 29 WP and EDPS and states that in order for consent to be unambiguous "a clear affirmative action"[79] is required[80].

[79] See Commission Staff Working Paper, Impact Assessment, Annex 2, p. 20 and also pp. 105-106: *"As also pointed out in the opinion adopted by WP29 on consent, it seems essential to clarify that valid consent requires the use of mechanisms that leave no doubt of the data subject's intention to consent, while making clear that – in the context of the on-line environment – the use of default options which the data subject is required to modify in order to reject the processing ('consent based on silence') does not in itself constitute unambiguous consent. This would give individuals more control over their own data, whenever processing is based on his/her consent. As regards impact on data controllers, this would not have a major impact as it solely clarifies and better spells out the implications of the current Directive in relation to the conditions for a valid and meaningful consent from the data subject. In particular, to the extent that 'explicit' consent would clarify – by replacing "unambiguous" – the modalities and quality of consent and that it is not intended to extend the cases and situations where (explicit) consent should be used as a ground for processing, the impact of this measure on data controllers is not expected to be major."*

[80] See Article 4(11) GDPR. "Consent of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her."

2.3 Consent vs. big data: current challenges

As mentioned above, new technologies such as mobile devices, location-based services, the internet of things, and the existence of ubiquitous sensors have brought into question the means of obtaining users' consent for the processing of their personal data.
the processing of their personal data.

The solution has been seen in online privacy policies, offered to users as unilateral and (quasi) contractual terms, which have become the cornerstone of online privacy protection, despite the overwhelming evidence that most people do not even read the terms or do not understand them.

Faced with this situation, legal practitioners are calling for improvements, especially with regard to:

- the way in which privacy policies are drafted, so that there is an effective notification;
- and developing mechanisms for granting informed consent, with particular emphasis on opt-in and opt-out systems.

For their part, the challenges arising from big data mean that consent, in itself, is not sufficient. We will now analyze the challenges faced by existing models of notification and consent.

(i) Is simple language the solution?

An ideal privacy policy would offer users real freedom of choice, based on a sufficient understanding of what that choice entails. Some stakeholders advocate the use of plain language, easy-to-understand policies, and easy-to-identify boxes or windows where users can indicate their consent.

However, in today's environment of complex data flows and stakeholders with different interests, what Solon Baroccas and Helen Nissebaum have called “the transparency paradox”[81] has been triggered, in the sense that simplicity and clarity inevitably lead to a loss of precision.

Baroccas and Nissebaum state that the evidence in this regard is stark: the few users who read privacy policies do not understand them. Thus, a simple wording of these privacy

[81] Barocas, Solon and Nissembaum, Helen. “Privacy, big data and the public good»*;* Chapter 2*:* Big data's End Run Around Anonimity and Consent”, *Cambrigde University Press* (2014).

policies could make them easier to understand. However, even when users do understand privacy policies, texts written in this simple language do not provide sufficient information for informed consent. On the other hand, the detail that would be necessary for the privacy policy to provide sufficient information would be overwhelming[82].

Thus, for example, in the personalized advertising business, which is highly developed in today's big data era, in order for users to make an informed privacy decision, they should be notified about the type of information that is collected, with whom it will be shared, under what limits, and for what purposes. Plain language cannot provide all the information necessary for users to make a sufficiently informed decision.

In this regard, it has been estimated that if all US internet users were to read privacy policies every time they visit a new web page, the country would lose about $781 billion annually due to the opportunity cost of the time spent reading these privacy policies[83].

(ii) Duty to report primary and secondary data

The same problems arise, not only with respect to personalized advertising but also in general. For example, let us consider some moments in which data is generated and stored on a daily basis: opening a profile on a social network, shopping on the Internet, downloading a mobile application, or traveling. All these activities create raw data whose further processing justifies the individual's consent.

Moreover, the chain of data senders and receivers is potentially infinite and includes actors and institutions whose roles and responsibilities are not delimited or understood. Thus, the transfer of data can become relatively obscure.

What has been said so far begs the question: how should the information be redacted so that users can give their informed consent? The way big data works makes this task tremendously difficult, as data moves from one place to another and from one recipient to

[82] Idem

[83] Aleecia M. McDonald and Lorrie Faith Cranor. "The Cost of Reading Privacy Policies". *Journal of Law and Policy of the Information Society,* Vol. 4, n.o 3 (2008).

another in an unpredictable way since the value of the data is not known at the time it is collected. Thus, consent increasingly resembles a blank check[84].

In this situation, the question that arises is whether the data controller's obligation to inform about the collection of data is limited to the information he or she explicitly collects, or whether a broader criterion should be adopted and this duty to inform should be understood as also extending to information that the institution may obtain after processing.

Many authors are of the opinion that the duty to inform and the need to obtain consent should refer not only to the fact that primary data are collected but also to the information that can be extracted from a sophisticated analysis of them, including the information that can be extracted from the aggregation of data collected by the company with data from other sources and files. However, this approach has many practical difficulties, since, by its very nature, the value of big data lies precisely in the unexpectedness of the results it reveals. So how does the data controller explain that it is impossible to know in advance what information the processing of the data collected will reveal? Many authors consider that the consent given under these circumstances is not the informed consent required by law[85].

(iii) Deriving majority data from minority data

The dilemma over consent to collect and process personal data is also compounded by what Baroccas and Nissebaum have denounced as the "tyranny of the minority". Their theory is based on the premise that information voluntarily shared by a few individuals can reveal the

[84] Barocas, Solon and Nissembaum, Helen.. "Privacy, big data and the public good»*;* Chapter 2*:* Big data's End Run Around Anonimity and Consent", *Cambrigde University Press* (2014).

[85] By all, Fred H. Cate and Viktor Mayer-Schönberger."Notice and consent in a world of Big data". *International Data Privacy Law,* Vol 3, n.o 2 (2013); Omer Tene and Jules Polonestsky. "Big data for all: Privacy and user control in the age on analytics". *Northwestern Journal of Technology and Intellectual Property*, Vol. 11, n.o 5 (2013); Ira S. Rubinstein "Big data: The End Of Privacy Or A New Beginning?" *International Privacy Law,* Vol. 3, n.o 2 (2013).

same amount of information about those who choose not to consent, while the institutions with whom that minority have consented can infer the same results for the majority who have not consented[86].

This implies that, in reality, each individual has no real ability to make a decision that protects his or her interests (in this case, to protect his or her privacy and the information that his or her data may reveal). Let us look at some practical examples.

For example, a friend request on a social network shows a connection between the two people, which allows inferring some kind of common link; either shared interests, affinities, or some personal history. This makes it possible to create inferences on certain behaviors.

Under this premise, computer scientists have set to work to try to answer the question of whether big data techniques of social network analysis and data mining could be used to infer attributes of a user based on information revealed by another user. The results of several of the experiments that have been carried out are revealing.

Example: Inferring information about a person from information provided by his or her friends on social networks

Social network users create profiles that normally include data such as geographic location, interests, and the university they attend. This information is used by social networks to group users, share content, and suggest connections with other users. But not all users disclose this information.

In his experiment, Alan Misolve[87] wanted to answer the following question: given certain attributes by some of the users of a social network, can we infer those same attributes about other users, using social relationship graphs?

To try to answer this question, the study collected very detailed data from two social networks. They observed that users with common characteristics are more likely to be friends on social networks, and sometimes create dense communities. The study showed that it is possible to infer attributes from communities and friendship relationships in social networks. Thus, the study analyzed the information that some social network users posted about their university degree, their year of graduation, and their dormitory in university dorms. The

[86] Barocas, Solon and Nissembaum, Helen.. "Privacy, big data and the public good»*;* Chapter 2*:* Big data's End Run Around Anonimity and Consent", *Cambrigde University Press* (2014).

[87] Alan Misolve et al. «You are who you know: inferring users profiles in online social networks». *Web Search and Data Mining (WSDM).* ACM, New York (2010).

experiment was able to deduce these same attributes, with a high degree of accuracy, from those other students who had not disclosed these data on social networks.

In doing so, the study concluded that some attributes can be inferred with a high degree of accuracy from the data of only 20% of the users.

Example: Inferring a person's sexual orientation from social network friendships

Two students at the Massachusetts Institute of Technology (MIT) created a software program called Gaydar in 2007 to infer the sexual orientation of social network users as a project for their Ethics, Law, and the Internet[88] course.

They analyzed the Facebook friendship ties of 1544 men who declared themselves heterosexual, 21 men who declared themselves bisexual, and 33 homosexuals, and investigated the correlations between the user's sexual orientation and that of his friends. Homosexual men had a much higher proportion of homosexual friends on the social network, so with this data, they created a system for the program to infer the sexual orientation of other users based on their friends.

The study could not prove its conclusions with scientific rigor, as it was restricted by the limitations of being a final-year project. However, it was able to demonstrate that the program could deduce with a small margin of error which users were homosexual men. In contrast, predictions about bisexual men or homosexual women were not as accurate.

Gaydar is just one of many projects that aim to data-mine users' social network information and friendship relationships to obtain potentially very valuable, but personal, information.

The risk to people's privacy is increased when, from the confirmed data of a sufficiently large number of social network contacts, who disclose their own data, undisclosed data of other users can be inferred. Because of this, certain people decide not to be active on social networks such as Facebook. However, this solution may not be sufficient.

[88] Carter Jernigan and Behran F. T. Mistree. «Gaydar: facebook friends- hips expose sexual orientation». *First Monday*, Vol. 14, n.o 10 (2009).

Example: inferring information about individuals who are not part of a social network from information obtained in social networks

The experiment carried out by a team from the University of Heidelberg (Germany) wanted to analyze whether it is possible for data revealed by certain individuals in social networks to yield data about other individuals who are not part of the social network[89]. By studying graphs of social relationships in networks and the e-mail addresses of network members, the group created an algorithm capable of inferring that two people outside the social network (and even without knowing each other) shared certain characteristics, based on data obtained from a common friend present in social networks.

In this regard, communication and social network researcher Danah Boyd states that "your permanent record is no longer just about what you do. Everything that others do that connects us, involves us or can influence us will become part of our permanent record"[90].

Thus, even if a user makes efforts not to reveal personal information (for example, by changing the computer's default settings, refusing to publish information about their political ideology, religion or sexual orientation, or not publishing photos), information about their contacts on social networks, or even the same list of contacts on social networks, can allow others to deduce information about us.

But even more astonishing is the fact that similar inferences can be made about an entire population even when only a small proportion of people, with whom they do not even have connections or friendships, reveal their data.

Example: Predicting pregnancies

The Target department store chain conducted a study through which it could predict the pregnancy rate of its customers. In this case, no inferences were made on the basis of friends on social networks. Target analyzed data from its files on women who had held a baby shower to identify women who had disclosed the fact that they were pregnant, and studied their shopping basket. Since these habits were different from those of other customers, Target was able to find out which customers might be pregnant on the basis of the change in their

[89] Emöke-Ágnes Horvát, Michael Hanselmann, Fred A. Hamprecht, and Katharina A. Zweig. "One plus one makes three (for Social Net- works)". *PLOS One Journal* (2012).
[90] Boyd, Danah. "Networked privacy". *Personal Democracy Forum*, New York (2011).

consumption habits and the information revealed by those other women with whom they had no ties.

Thus, the question arises: What is the minimum proportion of people who must disclose their data on a particular attribute in order for it to be possible to identify which other members of the total population possess that same attribute?

This question brings us back to statistical concepts: as long as the sample is representative and the attributes analyzed are statistically relevant, it will be possible to make inferences with a smaller margin of error from a smaller sample.

In this sense, the study by Alan Misolve and colleagues[91], to which we referred a few pages back, revealed that it is possible to infer certain attributes from the entire population if only 20% of this population reveals these attributes. It should be noted that this threshold was obtained for this particular experiment, which sought to ascertain relatively simple attributes (degree studied, year of graduation, and dormitory), and only analyzed the very attributes that we then wanted to infer, without assessing the other large amount of information that can be extracted from the social networks.

In any case, it seems reasonable to conclude that the added value of the consent given by a specific individual decreases as other users give their consent and the sample and the database that is created become statistically representative.

It is in this way that a minority can determine the attributes to be inferred from the total population analyzed, and it discourages data controllers from investing in processes that facilitate obtaining consent from the remaining users once the minimum threshold of representativeness has been reached. In sequence, not providing consent may not change the way data controllers categorize or treat an individual.

(iv) Loss of social benefit and innovation

[91] Alan Misolve et al. «You are who you know: inferring users profiles in online social networks». *Web Search and Data Mining (WSDM).* ACM, New York

As stated at the beginning of this chapter, Article 4(11) of the GDPR emphasizes that consent "seems to imply a need for action". In practice, this implies prioritizing opt-in systems over opt-out systems.

Opt-in and opt-out systems are two ways of manifesting consent. The opt-in system is based on the user's express and positive consent by filling in the box created for this purpose. In the opt-out system, on the other hand, the individual must express his or her opposition, either by filling in the corresponding box or by informing the organization by the appropriate means (for example, on many occasions the way to express opposition is by accessing a web link to unsubscribe)[92].

In summary, as we already know, consent is the main instrument of data protection regulations, the basis of which is to give the individual power of control over his or her data. Opt-in and opt-out systems allow individuals a different degree of control over their data. And is this difference relevant in practice?

To analyze it, let us first take a break to point out two currently existing doctrinal lines. On the one hand, some authors[93] argue that all data should be considered personal, and thus subject to the requirements of data protection regulations. This would imply requesting consent for the processing of any data. However, such a broad definition of personal data would be factually unmanageable.

In contrast to this trend, other authors[94] argue that a more pragmatic approach to individual consent should take precedence and that the right to privacy and data protection should be balanced with other social values such as public health or national security. Thus, in relation to the opt-in and opt-out systems mentioned above, they consider that the opt-in system could lead to a significant loss of collective social benefits if individuals choose not to fill in the box to give their consent.

In an attempt to propose a solution to the differences between these two currents, Tene and Polonetsky propose a framework based on a risk matrix: when the benefits of data processing outweigh the risks to the privacy of individuals, it should be assumed that the data controller has the legitimacy to process the data, even when individuals have not consented[95]. For example, analyzing websites to understand and improve the use of that website creates high

[92] Bellman, Steven & Johnson, Eric & Lohse, Gerald. (2001). To Opt-in or Opt-out? It Depends on the Question. Commun. ACM. 44. 25-27.
[93] Paul Ohm. “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization”. *UCLA Law Review*, Vol. 57 (2010).
[94] Omer Tene and Jules Polonetsky. “Privacy in the age of big data: a time for big decisions”. *Standford Law Review Online*, Vol. 64, n.o 63 (2012).
[95] Idem.

value, ensuring that products and services can be improved to achieve better service. The risks to privacy are minimal because if the system is implemented correctly, it only deals with statistical data that does not allow the identification of a specific individual. In these situations, requiring users to opt-in to the analysis would seriously threaten the use of the analysis. In fact, one of the keys to big data is to be able to differentiate individuals in order to keep each subject's information distinct from that of other subjects, but without having to identify them.

Let us also think, for example, of organ donation systems. In countries with an opt-in system, the rate of organ donation is much lower than in culturally similar countries with an opt-out system[96]. Thus, the donation rate is much higher in Sweden, which follows an opt-out model, than in Denmark, which follows an opt-in model, despite the fact that they are culturally very similar countries and behave similarly in many other areas. The same is true of Austria and Germany.

Tene and Polonetsky do not argue that express consent should never be sought from the individual. In many instances, consent will be required (either through an opt-in or opt-out system), such as for behavioral marketing services, third-party data processing, or geolocation-based services[97]. However, an increasing focus on opt-in consent and the principle of data minimization, without taking into consideration the value or uses of such data, could slow down innovation and social advances.

The previous pages have shown that consent, as it is currently envisaged, does not solve the practical problems that it used to solve. Privacy policies are not understandable to individuals, and the information needed to comply with the legal requirement of informed consent cannot be detailed at the time consent is sought, precisely because the purposes for which the data may be used are not known a priori.

In the following pages, we turn to the other major instrument that makes it possible to obtain the value of the data while preserving the privacy of the individuals: the techniques by which the data are made anonymous.

[96] Idem.
[97] Idem.

Chapter 3: Anonymization and Pseudonymization of Data

At this point, we already know that there is a dichotomy between personal data (which are subject to data protection rules) and anonymized data. Once a dataset has been anonymized and individuals are not identifiable, data protection regulations do not apply.

Traditionally, anonymization consisted of a two-step process. First, data sets were stripped of all personally identifiable information (PII), such as name, address, date of birth, or social security number. Second, other categories of data that could act as identifiers in that particular context were modified or removed (e.g., a bank would remove credit card numbers, and a university would remove student ID numbers).

Thus, the result combined the best of both worlds: the data remained useful, and could be analyzed, shared, or made available to the public while individuals could not be identified, thus protecting their privacy. Anonymization ensured privacy.

However, with new developments, this situation is changing. Big data, by increasing the quantity and diversity of information, facilitates the re-identification of individuals, even after they have been anonymized[98].

Indeed, practice shows that creating a truly anonymized dataset is no easy task. For example, a set that is considered anonymous may be combined with others in such a way that one or more individuals can be re-identified. This is why anonymization is a critical process for data protection authorities.

In this regard, the US Federal Trade Commission has stated that:

"There is sufficient evidence to show that technological advances and the ability to combine different data can lead to the identification of a consumer's computer or device, even if this data by itself does not constitute personally identifiable data. Moreover, not only is it possible to re-identify non-personally identifiable data through various means, but companies have strong incentives to do so"[99].

[98] Kenneth Neil Cukier y Viktor Mayer-Schoenberger. «Big data: A Revolution That Will Transform How We Live, Work And Think». *Houghton Mifflin Harcourt* (2013).
[99] Federal Trade Commission (FTC). «Protecting Consumer Privacy in an Era of Rapid Change. Recommendations for Businesses and Policymakers» (2012).

3.1 Legal framework for anonymization

The data privacy concerns are not new. Already in January 1981, the European Council[100] adopted the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data. Since then, the Convention, known as Convention 108, has been extended to include issues such as protection in social networks, profiling, or the workplace. Since 2006, 28 January has been celebrated as Data Protection Day in Europe to commemorate the signing of the Convention. However, despite the existence of the Convention, the member countries did not have harmonized regulations in this area.

In 2000, the Charter of Fundamental Rights[101] of the EU included in its Article 8 the "Protection of personal data":

1. Everyone has the right to the protection of personal data concerning him or her.
2. Such data shall be processed fairly for specified purposes and on the basis of the consent of the data subject or on another legitimate basis laid down by law. Every person has the right of access to data concerning him or her which has been collected and the right to have it rectified.
3. Compliance with these rules shall be subject to control by an independent authority.

Subsequently, with the entry into force of the Lisbon Treaty[102] in December 2009, the EU Charter of Fundamental Rights became legally binding and, with it, the right to the protection of personal data was elevated to the status of a fundamental right independent of the right to privacy. In other words:

"The guarantee of a person's privacy and reputation now has a positive dimension that goes beyond the scope of the fundamental right to privacy and is translated into the right of control over personal data".

---

[100] Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data. (1981) https://rm.coe.int/1680078b37

[101] European Union, 2000 Charter of Fundamental Rights of the EU. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012P/TXT&from=EN

[102] Treaty of Lisbon in December 2009. https://www.europarl.europa.eu/ftu/pdf/en/FTU_1.1.5.pdf

The European Council's Europe 2020 strategy[103] is planned around 7 pillars, one of which is the Digital Agenda for Europe that promotes the creation of a free and secure European Digital Single Market[104] where businesses can sell across the EU and citizens can shop online across borders. The strategy for a digital single market was adopted in May 2015.

As part of this strategy, the harmonization of the aforementioned regulations, including personal data privacy, has been promoted, with the aim of creating a framework of trust so that an internal digital market can develop with legal certainty for users and transparency. These efforts culminated in the new General Data Protection Regulation[105], which entered into force as of 25 May 2018. This regulation constitutes a legal framework that will be applied directly in all the Member States of the EU, and that will have priority over national legislation.

But what is the purpose of the GDPR? The protection of personal data, that is, of information relating to natural persons. It is one of the largest restructurings in relation to the processing of personal data, which may affect not only companies but any person, entity, public authority, service, or other body that processes personal data of those who reside in the European Union. And this also includes suppliers and third-party companies that are entrusted with the processing of that personal data.

This regulation has a very broad scope of application since, in addition to affecting the EU countries, some such as the United Kingdom will also have to adapt since, despite leaving the European Union in 2019 (Brexit), the GDPR will be incorporated into British legislation. On the other hand, the GDPR also affects companies from outside the EU that offer goods or services to people from the European Union or that monitor their behavior within the EU. For example, it directly affects US companies that host websites accessible to people from the EU. It, therefore, has an unprecedented scope.

In reality, European legislation does not regulate anonymous data or the process of anonymizing information. According to the Directive, anonymization involves the processing of personal data in such a way that it is no longer possible to re-identify it:

---

[103] European Union, Europe 2020 Strategy of the European Council. https://ec.europa.eu/eu2020/pdf/COMPLET%20EN%20BARROSO%20%20%20007%20-%20Europe%202020%20-%20EN%20version.pdf

[104] European Union European Commission, European Digital Single Market. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52015DC0192&from=EN

[105] European Union, General Data Protection Regulation.

"The principles of protection shall apply to any information relating to an identified or identifiable person. In order to determine whether a person is identifiable, it is necessary to consider all the means that can reasonably be used by the controller or by any other person to identify that person, and the principles of protection shall not apply to data rendered anonymous in such a way that it is no longer possible to identify the data subject (...)"[106].

The Privacy and Electronic Communications Directive[107] has referred to the concepts of "anonymization" and "anonymized data" in similar terms. Thus, from the text of the Data Protection Directive the key elements of anonymization can be extracted:

- "All the means that can reasonably be used must be considered": this implies taking into account the state of the art at any given time, with particular reference to the cost and know-how required to reverse anonymization.

It could be criticized that the wording of the standard uses such abstract terms. However, we must bear in mind that the development of technology over time could increase the risk of re-identification of data so that the legal wording can be adapted to the context of the time, it must remain technically neutral.

- "Anonymization must be carried out in such a way that it is no longer possible to identify the data subject": should this be interpreted as meaning that the Directive sets the threshold for anonymization to be irreversible? We will answer this question in the following pages.

[106] Recital 26 of Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.
[107] Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications); also known as the "e-Privacy Directive".

3.2 At what threshold do we consider data to be anonymous?

The Directive has been transposed into the national legislation of each Member State and interpreted by its courts in very different ways. Thus, what is meant by anonymous data may vary between:

- Absolute anonymization: implies zero possibility of re-identifying, directly or indirectly, any of the subjects. In fact, this level of anonymization is often impossible to achieve, especially when dealing with very data-rich files.
- Functional anonymization: implies a negligible risk of re-identification.

On this point, the Article 29 WP opinion on anonymization techniques seems to fall into some contradictions that are important to analyze. On the one hand, the recital recognizes that there is a residual re-identification risk even after applying anonymization techniques. But, on the other hand, the recital also notes that the Directive mandates that anonymization be "irreversible." It seems that the two concepts are contrary to each other. The importance in the practice of assuming one or the other interpretation justifies the analysis.

In this regard, the work of Khaled El Emam and Cecilia Alvarez is of great help[108]. If the threshold for considering data to be anonymized is "zero risk", anonymization would not be possible in practice, and therefore, it would not be possible to process or assign data without consent (or one of the other legal grounds). This could lead to the cessation of many information flows, with the loss of social benefit that this would entail, even more so in the big data environment. It could also mean that organizations would have no incentive to make it difficult to identify data and would look for alternatives to continue processing data, which could include expanding the primary purposes of the data, which could ultimately be more harmful to the privacy of individuals.

Thus, the most reasonable interpretation is that "zero risk" should not be the threshold to be followed. Instead, the reasonableness test set out in recital 26 of the Directive, cited above, should be applied.

[108] Khaled El Emam and Cecilia Alvarez. “A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques”. *International Data Privacy Law Journal*, Oxford University Press, Vol. 5, n.o 1 (2015).

GDPR has also adopted this criterion, which "helps to achieve a uniform interpretation" in all Member States. And, as K. El Emam and C. Alvarez point out, this is also the definition developed by non-European jurisdictions, such as Canada[109] and the United States[110]. While absolute anonymization does not exist, Professor Mark Elliot refers to the concept of "re-identification risk-utility"[111]. That is, in order for the data collected to retain its usefulness, it operates under a concept of negligibility. Accordingly, when the re-identification risk is so small that, despite not being statistically zero, it is considered functionally zero, the anonymization carried out is considered sufficient. Otherwise, operating with the data available to us would be factually impossible.

In a similar vein, Professor Josep Domingo Ferrer clarifies how anonymization is carried out in 95% of cases: by means of the approach he calls "utility first". It consists of anonymizing the data while maintaining its utility and then analyzing the level of risk of re-identification. If the risk is too high, the data should be anonymized again with a higher distortion until the risk analysis yields an acceptably low level[112].

All these approaches by K. El Emam, C. Alvarez[113], Mark Elliot[114], and Josep Domingo Ferrer[115] are similar in that they try to preserve the usefulness of the data.

In contrast to this approach is what Professor Ferrer calls "privacy-first", which has been adopted mainly in the academic world, but not often used in practice. This method consists of establishing a priori the desired level of privacy and using privacy models that guarantee this level of protection, without taking into account the level of utility of the data[116].

Thus, the central question is what should be considered an acceptable level of re-identification risk, while anonymized data fall outside the scope of data protection regulations.

In the context of big data, the analysis of large amounts of data may make it possible to identify individuals from data that no one would have considered personally identifiable or

---

[109] Personal Health Information Protection Act, Ontario, Canada (2004).

[110] Health Insurance Portability and Accountability Act Privacy Rule (HIPAA), the United States (1996).

[111] Mark Elliot. "To be or not to be (anonymous)? Anonymity in the age of big and open data". [Conference] *Computers, Privacy* & *Data Protection on the Move (CPDP)* (2015).

[112] Domingo-Ferrer, Josep. "Big Data Anonymization Requirements vs Privacy Models". *471-478.* (2018).

[113] Khaled El Emam and Cecilia Alvarez. "A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques". *International Data Privacy Law Journal*, Oxford University Press, Vol. 5, n.o 1 (2015).

[114] Mark Elliot. "To be or not to be (anonymous)? Anonymity in the age of big and open data". [Conference] *Computers, Privacy* & *Data Protection on the Move (CPDP)* (2015).

[115] Domingo-Ferrer, Josep. "Big Data Anonymization Requirements vs Privacy Models". *471-478.* (2018).

[116] Idem.

previously anonymous, as well as to infer the tastes and behaviors of individuals despite not knowing their identity.

3.3 Pseudonymization is not anonymization

Traditionally, pseudonymization was included as an additional anonymization technique. Pseudonymization consists of replacing one attribute of a dataset (normally a unique attribute that acts as a direct identifier, such as the first and last name) with another attribute (such as, for example, the DNI, the Social Security number, or an alias code that cannot be deciphered, so that it cannot be known to whom it refers)[117].

The most widespread methods of pseudonymization are encryption and tokenization.

Encryption and tokenization. Definition

Encryption with a secret key allows the owner of the key to re-identify the subjects by decrypting the key (for example, by re-associating each Social Security number with the person's name).

Tokenization, on the other hand, is mainly applied in the financial sector for credit card processing. Normally, the creation of the identifier (token) consists of replacing the ID card numbers with values of little use to a potential hacker, but which guarantee the same operability, through a one-way encryption system that generates a random number.

Thus, although pseudonymized data were traditionally considered anonymous data, today pseudonymization is no longer considered a method of anonymization, since the person is still identifiable, albeit indirectly.

Insofar as it reduces the linkability between the information and the subject from which it originates, pseudonymization is a useful security measure, although it still allows the identification of subjects. Thus, pseudonymized data are currently considered to still be personal data and are subject to the regulations on the protection of personal data[118]. In this respect, the history of the Social Security number is illustrative. By associating each person with a unique number, this would be the person's identifier in the eyes of the administrations,

[117] Article 29 Working Party. "Opinion 05/2014 on Anonymisation techniques" (2014).

[118] Mark WIliot. "To be or not to be (anonymous)? Anonymity in the age of big and open data". [Conference] *Computers, Privacy* & *Data Pro- tection on the Move (CPDP)* (2015); Khaled El Amam and Cecilia Alvarez. "A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques". *International Data Privacy Law Journal*, Oxford University Press, Vol. 5, n.o 1 (2015).

although outside these it was a meaningless, anonymous piece of data. In fact, the number was even a more unique identifier than a name, since the most common ones can appear more than once in a database. As its use became more widespread, large companies first, and then small businesses, adopted it as a way of identifying individuals, and it is now a unique and commonly used identifier. Therefore, the social security number is currently sensitive data that allows direct identification of the person and is, therefore, personal data.

Thus, any random data that can be used as a unique identifier will be a pseudonym and not an "anonymous identifier", whose anonymous value decreases as its use becomes more widespread.

This does not imply that the use of pseudonyms is worthless. Indeed, it limits the ability to infer data such as gender, race, religion, or national origin. Moreover, those pseudonyms that are not shared between different databases do not allow such direct identification. In principle, only the institution that assigns this pseudonym will be able to recognize the person to whom it corresponds. And when the pseudonym is removed or replaced, not even the institution that assigned it to a particular person will be able to recognize that person by the same pseudonym[119]. This is what happens, for example, when a cookie that we have installed on our computer and that allows us to identify ourselves is deleted from the system when it expires or when we delete it.

This being the case, pseudonymization cannot currently be considered an anonymization technique. However, the mistake of considering it to be an anonymization method is one of the most important risks of pseudonymization techniques.

Particularly revealing in this regard is the work of Narayanan and Shmatikov, who have shown the effect of re-identification in the context of social networks[120].

Case: friendship graphs in social networks

[119] Barocas, Solon and Nissembaum, Helen. "Privacy, big data and the public good. Chapter 2: Big data's End Run Around Anonimity And Consent". *Cambrigde University Press* (2014).
[120] Arvind Narayanan and Vitaly Shmatikov. "De-anonymizing social networks*". The University of Texas at Austin, Simposio IEEE on Security and Privacy* (2009).

Narayanan and Shmatikov have conducted a study in which they have shown that certain sensitive information about social network users can be extracted from social network graphs, even though the data have been pseudonymized by using nicknames instead of the subjects' real names. This is possible because the friendship relationships between individuals in a social network are unique and can serve as identifiers.

The experiment consisted of creating an algorithm that made it possible to re-identify the users of the Twitter social network through the friendship graphs of the Twitter network itself and of the Flickr social network, which was used as a source of additional information.

Thus, the first graph was formed by the follower relationships of the Twitter network. The second graph was formed by the Flickr contact network. Both networks require the mandatory use of a user name and also allow the additional fields of name and location to be filled in. The relationship graphs use only the usernames (e.g. "xk_562"), so they are "anonymous" (pseudonyms in fact).

The experiment also used graphs showing friendship relationships from the LiveJournal blog network, to which the researchers had access.

To create the algorithm, Narayanan and Shmatikov[121] tested the known data from the LiveJournal network to see how much auxiliary information is needed to identify a node in the target social network.

The algorithm then carries out what is called a "re-identification attack" in two stages. First, the attacker identifies the nodes of the friendship graphs that are present in both social networks, i.e. (i) the targeted network, Twitter, and (ii) the network whose graph serves as auxiliary information, Flickr[122]. Subsequently, both graphs are superimposed. The information obtained by overlaying both graphs creates new data that feed the algorithm, which in turn identifies new information to expand the friendship graphs. The result is a superposition of the graphs of the auxiliary network and the target network in such a way that the subjects of these accounts can be reidentified. The result of the study is that up to one-third of the individuals (with verified accounts) who are members of both social networks, Twitter and Flickr, can be re-identified with only a 12% error rate. The data that was taken to conduct the experiment is

[121] Idem.
[122] Idem.

more restricted than the data that a real attacker could obtain, so the hit rate in the experiment is lower than what the real attacker would achieve.

Also, the number of individuals present in both social networks at the time of the experiment was not very high (less than 15%), so the initial graph overlap was not very large. However, other larger social networks have a higher overlap ratio (for example, the study itself indicates that about 64% of the users of the Facebook network at that time were also users of the MySpace social network). All this implies that this algorithm can achieve a much higher re-identification rate in other social networks.

As they state in their conclusions, "the main lesson of this work is that anonymization - which is actually pseudonymization - is not enough to achieve privacy when it comes to social networks. We have developed a generic re-identification algorithm and have shown that several thousand users can be successfully de-anonymized in the anonymous network graph of the popular microblogging network Twitter, using a completely different social network (Flickr) as an auxiliary source of information"[123].

A real-world example of the problems that can arise when pseudonymization is considered sufficient to achieve anonymization is the well-known incident suffered by America On-Line (AOL) in 2006[124].

AOL case:

In 2006, the Internet service provider, AOL made public the data of 20 million Internet searches that users had performed on its engine, corresponding to 657,000 users, with the sole idea of favoring free research. The data had been "anonymized" (actually pseudonymized), removing the user name and IP address and replacing the data with unique numeric IDs that allowed researchers to correlate different searches with individual users. The goal was for

[123] Idem.
[124] Paul Ohm. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization". *UCLA Law Review*, Vol. 57 (2010).

researchers to be able to link searches conducted by the same person, without accessing personal information.

Within days, certain searches were linked, such as those of the user assigned the code 17556639: "how to kill your wife", "car accident photo", "photos of dead people", although in this case the subject was not identified.

A few days later, however, the New York Times did manage to identify Thelma Arnold, a 62-year-old widowed woman from the town of Lilburn, through its searches.

As Article 29 WP has stated, Internet search history, coupled with other attributes such as a customer's IP address, or other customer information, has a high power of identification[125].

If the incident had taken place on European territory, the legal reasoning would be as follows: as long as the data were merely pseudonymized or disassociated, they were still subject to the Data Protection Directive. Therefore, the processing of the data had to be compatible with the purpose for which it was collected, which of course did not include publication or attempted re-identification. The security flaw was caused by considering that pseudonymization was the same as anonymization.

In both cases, the adversaries needed an external data source to rejoin the data with their identifier. It is clear from these cases that, as far as privacy is concerned, there is a particular preoccupation with first and last names as personal identifiers, but in reality, any sufficiently unique pattern can be used to recognize the same person in other databases[126].

[125] Article 29 Working Party. "Opinion 05/2014 on Anonymisation techniques" (2014).

[126] Barocas, Solon and Nissembaum, Helen. "Privacy, big data and the public good. Chapter 2: Big data's End Run Around Anonimity And Consent". *Cambrigde University Press* (2014).

3.4 Criticism of the anonymization criterion proposed by the article 29 Working Party

The WP 29 Opinion discusses different anonymization techniques, with varying levels of robustness. As none of these techniques eliminates the risk of re-identification completely, it will normally be necessary to combine different techniques.

In the aforementioned opinion, the Working Party proposes two methods for determining whether a file is anonymous:

1. Conduct an analysis on the risk of re-identification of the data.
2. Verify that the file does not have any of the following properties:

I. Ability to single out an individual ("singling out"), which is defined as the ability to isolate some or all of the data of the same individual within the file.

II. The ability to associate at least two pieces of data belonging to the same subject or group of subjects, either in the same file or in two different files ("linkability"). Article 29 WP goes on to explain that the difference between this attribute and the previous one is that if an adversary is able to determine that a piece of data refers to the same group of individuals, but cannot single out the individuals in that group, an association will have occurred, but not a single out.

III. Ability to infer new data about individuals ("inference"): that is, to be able to deduce, with a significantly high probability, a new datum from the supposedly "anonymized" data.

These methods do not derive from the Directive, and would also result in reducing the usefulness of the data in an extreme way, so it seems important to dwell on this point.

The Art. 29 WP criterion of requiring that an anonymized file should not make it possible to single out a specific individual from the data in the file seems reasonable. Certainly, this property would imply that an individual is identifiable through the data contained in the file and that such data would therefore constitute personal data, and would in no case be anonymous data.

On the other hand, in my opinion, it is necessary to qualify the reasoning of the Art. 29 WP with regard to not allowing a file to have the properties of association and inference.

(i) The association of data pertaining to the same individual

The Art. 29 WP Opinion considers linkability to be negative, meaning the ability to associate at least two pieces of information belonging to the same subject or group of subjects, either in the same file or in two different files.

First of all, it is worth mentioning that information that makes it possible to identify a group of subjects, but not a particular individual within that group is not personal information. This follows from the very definition of personal data and has been assumed by other data protection institutions, such as the English Information Commissioner's Office (ICO)[127].

Khaled El Emam and Cecilia Álvarez point out that being able to link different data on the same individual within the same file is essential for creating longitudinal data sets. Thus, on this point, they oppose the Art. 29 WP criterion and emphasize that there are effective methods of anonymizing longitudinal data, so it is unreasonable to prevent their creation[128].

Longitudinal data. Definition

In very succinct terms, longitudinal studies are those in which repeated or follow-up measurements are made on individuals[129]. This makes it possible to study a group of individuals repeatedly over the years.

Longitudinal data are widely used in demographic and medical research. Some of their objectives are, for example, to describe the evolution of a patient, either before or after starting treatment or to make predictions of certain diseases[130].

Indeed, being able to associate, for example, all the symptoms that an individual shows at medical check-ups over the years is the basis for creating longitudinal data. To do this, it is necessary to be able to associate that same individual within the file, and all his or her associated data.

---

[127] Information commissioner's office (ICO, UK data protection authority). "Anonymisation: managing data protection risk code of practice" (2012).

[128] Khaled El Amam and Cecilia Alvarez. "A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques". *International Data Privacy Law Journal*, Oxford University Press, Vol. 5, n.o 1 (2015).

[129] Miguel Delgado Rodríguez and Javier Llorca Díaz. "Longitudinal studies: concept and characteristics". Spanish magazine of Public Health, Vol. 78, no. 2 (2004).

[130] Domingo-Salvany, Antonia and Brugal Puig, Maria Teresa and Barrio Anta, Gregorio. "SET Treaty of Addictive Disorders". Spanish Society of Addictions. Editorial Médica Panamericana (2006).

The Art. 29 WP Opinion also mentions a problem with the pseudonymization method that it does not remove the association property of the individuals' data. But in fact, as K. El Emam and C. Alvarez point out, precisely one of the main virtues of pseudonymous data is that it allows linking or associating data belonging to the same subject without the need to use personal identifiers. However, as we have already mentioned, pseudonymization is not a method of anonymization. The Information Commissioner's Office (ICO) also stresses the importance of maintaining the property of associating data to maintain its usefulness, in particular, for conducting longitudinal studies:

"Although pseudonymized data may not identify an individual in the hands of those who do not hold the "key", the ability to associate multiple databases to the same individual may be a precursor to the identification. In any case, this does not mean that effective anonymization through pseudonymization is impossible. The Information Commission (ICO) recognizes that certain types of research, e.g., longitudinal studies, can only be developed when different data can be reliably associated with the same individual (...)"[131].

In other words, the ICO recognizes the risk to privacy posed by associating data held in different databases. However, it can be understood that when they are in the same database, maintaining the ability to associate the data allows new knowledge to be obtained that can be used in a beneficial way.

[131] Information commissioner's office (ICO). «Anonymisation: managing data protection risk code of practice» (2012).

(ii) Inference

Similar to the association property, Art. 29 WP also considers the ability to make inferences about the data to be negative. The Opinion defines inference as the ability to infer, with a significant level of probability, the value of an attribute from the values of a set of other attributes.

Restricting the ability to make inferences has enormous consequences for big data. Let us dwell once again on statistics to understand the concept of inference and its importance. Statistical inference methods can be divided into two: parameter estimation methods and hypothesis testing methods. Well, when we estimate a parameter, an estimation error is made, which is the difference between our estimate and the real value of the parameter. With our estimate and this error, we construct a confidence interval, which is the probability that our estimate contains the true value of the parameter.

Let us recall the characteristics with which we defined big data: volume, variety, velocity. The enormous variety of data sources, and the speed at which they are created, mean that the data that serve as the basis for big data are often chaotic, incomplete, noisy, or unrepresentative. This poses a challenge for statistics, which can be overcome in part because, thanks to the greater volume of data that big data allows us to process, the estimates we can make are more accurate. This is because a model is simply a representation of reality. There are aspects of the model that are the same as reality, but other aspects are not found in the model. Big data allows the model to be better represented by having a larger volume of data. In any case, all models have a certain degree of inaccuracy or margin of error.

For their part, big data technologies make use of data mining to search for correlations in the stored data. Prof. Aluja uses the definition of data mining already advanced by Hans (1998): "the process of secondary analysis of large databases with the aim of finding unsuspected relationships that are of interest or provide value to the database holder[132]".

Data mining is, in fact, an evolution of data analysis statistics, which makes use of the integration of algorithms and process automation. Thus, one of its most important applications is to search for associations between events. That is, to answer the question: can we infer that

[132] Aluja, Tomas. "Data mining, between statistics and Artificial Intelligence". Polytechnic University of Catalonia, Vol. 25, n.o 3 (2001).

certain events occur simultaneously more than would be expected if they were independent?[133] This process has applications in an infinite number of fields, from marketing (for example, knowing that a consumer has bought product X, could he be interested in product Y?) to Genome Project research (for example, what gene sequences motivate the appearance of diseases?)

What some see as the greatest threat to privacy is, ironically, what makes big data most interesting: its ability to detect hidden correlations between data and thus infer conclusions that are not obvious to the naked eye. Indeed, one of the main uses of big data is to be able to make secondary use of the data to make inferences to obtain new knowledge.

However, inference can give rise to two types of new data. This is the basis of the argument of K. El Eman and C. Alvarez when they state that the Art. 29 WP Opinion treats two types of inference together, even though they do not normally go together in practice: the disclosure of a subject's identity and the disclosure of an attribute. Specifically, Art. 29 WP states that:

"It should be clear that "identification" does not only mean the possibility of retrieving a person's name or address, but also includes potential identification by singularization, association, and inference."[134]

On the one hand, the revelation of a subject's identity occurs when an adversary is able to assign a correct identity to a piece of information. This is what we mean when we speak of identification. Thus, an anonymous database is one in which the probability of inferring the identity of a subject is very low. On the other hand, there are attribute disclosures. That is when an adversary is able to learn something new about the subjects thanks to the analysis of the data. When we are faced with multiple variables, this is a complex process that is carried out by means of data mining and machine learning. An example might be to build a model from variables such as age, sex, previous patient diagnoses, and symptoms, to predict the probability of suffering from a certain type of cancer[135].

In this sense, Brian Dalessandro already stated that:

---

[133] Idem.

[134] Article 29 Working Party. "Opinion 05/2014 on Anonymisation techniques" (2014).

[135] Khaled El Amam and Cecilia Alvarez. "A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques". *International Data Privacy Law Journal*, Oxford University Press, Vol. 5, n.o 1 (2015).

"A great deal can be predicted about a person's actions without needing to know anything personal about them"[136].

This is a statement of enormous significance in the age of massive data. Conclusions obtained through data analysis with big data techniques can show new facts about individuals despite having no knowledge of their identity.

Instead of cross-referencing data entries associated with the same name or other personally identifiable information, data mining yields conclusions that allow companies to simply identify these characteristics. This opens the door for the qualities that can be inferred to go far beyond the information residing in the databases.

This also explains the statements made a few years ago by a Google engineer:

"We don't want the name. The name is noise. There is enough information in Google's huge databases about Internet searches, location and online behavior that you can find out a lot about a person indirectly"[137].

Having laid the groundwork for how the inference process works, the next question arises: is big data compatible with privacy? And if so, what are the legal implications of the greater ease of associating behavioral patterns and other data with the same individual despite not knowing his or her identity?

To answer this question I will go back to the legal definition above:

"the principles of protection shall apply to any information relating to an identified or identifiable person (...)"[138]

The rule seems to establish the requirement that identifiable data on "persons" be subject to data protection law. However, nothing is said about the possibility of making inferences about attributes (consumption trends, tastes, etc.). Thus, if big data or data mining techniques are used simply to build a model that allows inferring new characteristics or behavioral patterns without linking them to the identity of a specific person, there is no processing of personal data.

On the other hand, and as already mentioned, Art. 29 WP proposed a second means of determining whether a file is anonymous: an analysis of the risk of re-identifying the data. In

[136] Brian Dalessandro. "The science of privacy" *Ad: Tech (blog),* (July 30, 2013).
[137] Quentin Hardy. "Rethinking privacy in an era of big data". New York Times (June 4, 2012).
[138] Recital 26 of Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

this regard, it is sufficient to point out that inferring data does not undermine anonymization, precisely because the risk analysis will detect the real possibility of identifying a subject or group of subjects within the dataset. Thus, following this approach, we will also be able to use big data without inferences from new data resulting in personal data that must be subject to regulation.

In short, the ability to infer new characteristics or patterns of behavior should not be seen as an aggravation to the privacy of individuals, and anonymization models should not aim to prevent such inferences.

What has been explained so far has corresponded to what we have defined as the first phase of big data; that is, a process that has consisted of taking our data, inserting it into algorithms, and building a model that allows us to make inferences. From this point, we enter the second phase of big data, in which the new knowledge can be used to make decisions. It is at this step that a major risk to people's privacy may arise.

Decisions can be made about groups of individuals, without knowing the identity of each person (for example, when the government implements a vaccination policy against a new virus), or about specific individuals (for example, when people at risk of infection receive personalized information about their lifestyle and precautionary measures to avoid contagion).

As K. El Eman and C. Alvarez point out, the model can be used to make beneficial decisions such as those described above, but also decisions involving negative consequences such as discrimination. For example, if a person has been estimated to have a high risk of suffering from a rare disease and is estimated to have a life expectancy of 40 years, this information can be used to deny the person a scholarship or expensive medical treatment[139].

In conclusion, as many authors argue, the problem is not the capacity to make inferences, nor the model that is created with the new knowledge. The problem is the use made of that model. What is considered an appropriate use at any given moment will depend on the socially accepted norms and patterns, and will therefore be a subjective question that will have to be analyzed on a case-by-case basis.

[139] Khaled El Amam and Cecilia Alvarez. "A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques". *International Data Privacy Law Journal*, Oxford University Press, Vol. 5, n.o 1 (2015).

The solution proposed by K. El Eman and C. Alvarez is therefore different from that proposed by the Art. 29 WP. The opinion of the Art. 29 WP includes measures based on algorithms whose ultimate aim is to distort the data in such a way that neither singularization, association nor inference is possible. However, in the light of the above, it is clear that this solution would reduce the usefulness of the data to a minimum, preventing us from obtaining all the benefits that the new knowledge brings us. The solution proposed by these authors is the application of a model of "privacy ethics"[140].

In practice, this is done by creating an ethical council within the organization responsible for the data, working independently. This council would be made up of a representative of the individuals about whom decisions are to be made, a privacy and ethics expert, a representative of the company, and a representative of the public brand. Among the criteria to be taken into account, some of those that Art. 29 WP has already proposed in other reports and works are proposed:

- the relationship between the purpose for which the data were collected and the purpose of the current model;

- the context in which the data were obtained and the expectations of the subjects;

- the nature of the data and their potential impact;

- the safeguards to be applied, such as the imposition of conditions on the use of the data when they are transferred to a third party.

In my opinion, this last criterion deserves special importance. In any case, it will be the combination of these measures that can guarantee much more precisely that the use made of the data is lawful and legitimate, so that the social benefits derived from this use can be obtained, while at the same time protecting the privacy of individuals.

Again using statistical terms, this situation is what game theory would define as a win-win situation.

Having analyzed the anonymization criteria used by Art. 29 WP, subjected them to criticism, and proposed an alternative solution, let us now turn to the anonymization techniques set out in the Art. 29 WP Opinion.

[140] Idem.

(iii) Anonymization techniques

The Art. 29 WP Opinion on anonymization states that, in general terms, there are two approaches to anonymization techniques: randomization and generalization.

Randomization includes the family of techniques that alter the veracity of the data with the aim of eliminating the strong association between the data and the individual to whom it refers. Thus, when the data are sufficiently uncertain, they cannot be associated with a specific person.

The most commonly used randomization techniques are noise addition and permutation. Noise addition consists of modifying the data so that they are less precise but maintaining the general distribution of the data. Thus, for example, if our database collects the heights of the individuals in a study accurately, noise addition could consist of modifying the data so that they are accurate to within ±5 centimeters.

Permutation, on the other hand, consists of exchanging the attributes of individuals, so that they would be artificially linked to other subjects. This technique also allows the distribution of values to be maintained, although the correlation between individuals will be modified.

The second family of anonymization techniques is a generalization. It consists of generalizing or diluting the attributes of the subjects, modifying their scale (for example, referring to a country instead of a city; or to monthly data instead of weekly data).

3.5 Risk of re-identification

Recall that the Art. 29 WP Opinion proposed two criteria for determining whether a database is anonymous or not: (i) verify that the file does not have the properties of uniqueness, association, and inference, and (ii) perform an analysis on the risk of re-identification of the data.

So far we have reviewed the first criterion, and now we will move on to analyze the second proposed criterion.

As we mentioned in the introduction to this chapter, the advent of technologies such as big data and data mining has raised serious doubts about the power of anonymization.

In fact, the promise of achieving absolute anonymization is impossible to fulfill, especially for two reasons. First, because even if data do not contain information that can be considered personally identifiable information (PII), sometimes such data are still capable of uniquely distinguishing a person so that they can be associated with a specific person. Thus, for example, when the data contain extremely rich information (such as geolocation data), a person can be identified. And secondly, because so-called re-identification attacks are becoming more and more frequent and easier.

In Ohm's words, "data can be useful or perfectly anonymous, but never both"[141]. This is not to say that no anonymization technique can protect the privacy of individuals, as some techniques are truly difficult to reverse. However, technology has advanced and researchers (among others) have more than sufficiently demonstrated that anonymization can no longer be considered a panacea for data protection and privacy.

Let us look at two of the most paradigmatic cases of how databases that had been anonymized succumbed to a re-identification attack, calling into question the techniques used.

GIC case: identification by the trio of zip code, date of birth, and sex

[141] Paul Ohm. «Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization». *UCLA Law Review*, Vol. 57 (2010).

In the mid-1990s in Massachusetts, the Group Insurance Commission (GIC) decided to make public data on the hospital visits of public servants so that researchers could analyze them and draw conclusions. The GIC first "anonymized" the data by removing explicit personal identifiers such as name, address, and Social Security number. Despite this, nearly 100 attributes of each patient and hospital remained in the data made public, including the zip code, date of birth, and sex of the individuals. At the time, the Governor of Massachusetts, William Weld, publicly assured that the privacy of individuals was assured.

Latanya Sweeney, director of the Harvard University Privacy Lab, conducted a study aimed at highlighting the limitations of privacy standards[142] and security measures in order to implement improvements through the use of stronger and more complex algorithms.

Sweeney requested a copy of the published data and began trying to reidentify the Governor's data. He knew that the Governor resided in the city of Cambridge, Massachusetts, a city of 54,000 residents and seven zip codes. In addition, by paying $20, he purchased the latest voter census for the city of Cambridge, which contained, among other data, the name, address, zip code, date of birth, and sex of each voter.

Combining both databases, the GIC data, and the electoral roll, Sweeney managed to identify Governor Weld without difficulty: only six people in Cambridge were born on the same day as the Governor, only three of these people were male, and only he lived in his zip code. Sweeney sent the Governor's medical records, which included diagnoses and prescriptions, to his office. But Sweeney did not stop there. He extended his analysis to conclude that 87.1% of U.S. resident citizens are identifiable by this trio of attributes: zip code, date of birth, and sex.

The conclusions of Sweeney's experiment on this trio of identifiers have been reanalyzed and updated. In a study conducted at Stanford University, it was found that in 2006 the proportion of people living in the United States who could be re-identified using only the triad of zip code, date of birth, and sex had dropped to 63.3%. However, the new study again showed that re-identification attacks are still easy to carry out. Moreover, it is necessary to take into account that the availability of information that is currently public is much greater than at the time the study was conducted, and re-identification techniques are more accurate, so it is

[142] Specifically, Sweeney's study refers to the U.S. HIPAA privacy rule, which establishes measures to protect the privacy of medical and health data; however, the findings are extrapolable to other jurisdictions.

easy to imagine that an attacker could have access to more data than the zip code, date of birth and sex of the people whose data is in the database.

Another of the best-known cases that highlighted the limitations of anonymization (in this case carried out through randomization) is the case of the Netflix Prize[143].

Netflix case:

Netflix.Inc is the world's largest provider of monthly multimedia flat rates for watching movies and TV series. In October 2006, the company launched the so-called Netflix Prize. The company made public 100 million movie records of 500,000 users and offered a reward to the one who managed to improve its movie recommendation service (which is based on movies that other users with similar tastes rated very highly). The data had been anonymized so that all personal identifiers were removed except the movie ratings and the date of the rating; in addition, the noise was added, so that the users' ratings were slightly increased or decreased.

As an external data source, the public Internet Movie Database (IMB), an online database that stores movie-related information, was used.

The starting question of the experiment was: how much does an adversary have to know about a Netflix subscriber in order to be able to identify his data in the database, and thus know his complete movie history? That is, in analytical terms, the study was based on calculating the size of the ancillary data needed to re-identify the supposedly anonymized subjects. In this sense, one might ask whether a Netflix subscriber really considers his or her history of watching movies to be private. Even if the answer were negative (which cannot be assumed), that would be so only to the extent that we do not understand the real consequences of re-identifying this data.

As the experiment demonstrated, the correlation found between the anonymized Netflix Prize database and the public IMB database allows us to learn sensitive, non-public information about a person, such as political preferences or sexual orientation.

[143] Arvind Narayanan and Vitaly Shmatikov. "Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)". *The University of Texas at Austin* (2008).

Indeed, one of the users, a lesbian mother who kept her sexual orientation a secret and lived in a very conservative region of the United States, was identified and sued the company under the pseudonym, Jane Doe.

Following the scandal, researchers at the University of Texas compared Netflix data with other public data on movie ratings. The study showed that a user who had rated as few as six obscure movies (albeit from the top 500 list) could be identified in 84% of cases. This proportion increased to 99% of the cases if, in addition, the date on which the films were rated was known. Thus, it was shown that the rating of the films created a personal imprint.

Prior to this point, no one would have defined film ratings as personally identifiable data.

In both cases, it was necessary to combine two databases containing partial data on individuals, which demonstrates the principle that, although one database appears anonymous, when compared with a second database, unique information on the subjects is found, and re-identification of the subjects becomes possible.

This unique information about each person has been called a "fingerprint", i.e., combinations of data values that a person does not share with anyone else in the database and that therefore make it possible to identify him or her.

(i) Digital fingerprint

This fingerprint has generated countless debates and controversies about what is personally identifiable information. However, the above examples highlight a problem that transcends this fact: personal fingerprints have been found in data that until then had not been considered personally identifiable.

In other words, even though no names appear in a database, patterns can be seen. And with these patterns, a person with sufficient analytical skills can obtain names.

Thus, when it comes to big data, 2+2 does not equal 4. The synergies formed by joining different databases make the result greater than the mere sum of its parts.

The Electronic Frontier Foundation's project entitled "How unique -and trackable- is your browser?" (analyzes the browser data that create the device's fingerprint to infer how unique it is, and how capable it is of singling us out from other users.

Using its tool "Cover Your Tracks", formerly known as Panopticlik[144], it is possible to access the information that our browser voluntarily gives up when we visit web pages. Testing with my personal computer, the results are as follows[145]:

Before I deleted my cookies from my computer, the browser is identifiable among 5.3 million users.

After deleting my cookies, my browser is unique among almost 2.7 million users. In other words, it seems that deleting cookies makes our computer less unique, and therefore more anonymous. However, it is very remarkable that, even without cookies installed (at least the visible cookies), my computer is still perfectly unique among so many users.

[144] Panopticlick. Project "How unique —and trackable— is your browser?" Available in: https://coveryourtracks.eff.org/

[145] The experiment was conducted purely as a guideline, dated May 25, 2015.

(ii) Reidentification test: who is the adversary?

Common parlance has accepted the term "adversary" to refer to the person or entity that attempts to carry out a re-identification process or attack. According to the Directive:

"(...) to determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly."[146].

The Art. 29 WP Opinion on anonymization techniques uses these same terms. It follows that the adversary who may carry out a re-identification attack may be either the data controller or any other person.

However, the way in which one interprets who that other person may be has important consequences. Again following K. El Eman and C. Alvarez[147], we will take the following example to highlight the possible inconsistencies that arise when using the concept of "any other person".

Example: the importance of contextualizing who the adversary is

Imagine a hospital that gives previously anonymized data to a pharmaceutical company, which in turn has high security measures in place, is audited, and follows best practices in terms of data handling. In this environment, the likelihood of re-identification of individuals, whether deliberate or accidental, is very low.

In parallel, we know that there is Professor Slocum (a fictitious character), known for her ability to re-identify medical databases and to publish the results. In this context, Professor Slocum is an adversary who fits the definition of "any other person" and who can carry out a reidentification attack.

A strict interpretation of the Directive would lead to the conclusion that when the hospital transfers anonymized data to the pharmaceutical company, the high risk posed by

[146] Recital 26 of GDPR on the protection of individuals with regard to the processing of personal data and on the free movement of such data. https://www.privacy-regulation.eu/en/recital-26-GDPR.htm

[147] Khaled El Amam and Cecilia Alvarez. "A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques". *International Data Privacy Law Journal*, Oxford University Press, Vol. 5, n.o 1 (2015).

adversaries such as Professor Slocum must be taken into account. This would lead to the use of anonymization techniques that distort the data sufficiently so that, should Professor Slocum have access to this file, she would be unable to perform re-identification. This, as we have already seen, would mean that the usefulness of the data would be greatly diminished.

However, in practice, the data is only transferred to a specific entity such as the pharmaceutical company. The security measures that protect these data and the company's good practices make the actual likelihood of Professor Slocum having access to these data extremely low.

If the hospital is required to anonymize the data on the assumption at all times that Professor Slocum will have access to them and without regard to the actual context of the session, the usefulness of the data would be greatly diminished. This in turn would detract from the pharmaceutical company's efforts to put in place strong security measures and effective safeguards.

Indeed, taking the context into account is a criterion that has been embraced by Art. 29 WP itself. El Eman and C. Alvarez, in my opinion, solve the practical problems while maintaining privacy protection. It consists of considering that "any other person" refers to any third party who is in the same context as the one who is to receive the data, and who is also a "motivated intruder"[148].

The ICO anonymization code defines this motivated intruder as "a person who starts without any prior knowledge but seeks to identify the subject from whose personal data the anonymized data has been obtained"[149].

It is assumed that this motivated adversary is a reasonably competent person, who has access to data sources such as the Internet, libraries, and all public documents, and who will use investigative techniques. However, he has no special technical skills as a hacker, no access to specialized equipment, and will not resort to illegal schemes such as data theft. In the ICO's opinion, this is a good standard, as it sets the bar above a relatively inexperienced person from the general public, but below a specialist person with extensive analytical means or prior knowledge of the person.

---

[148] Idem.

[149] Information Commissioner's Office. "Anonymisation: managing data protection risk code of practice" (2012).

In my opinion, however, perhaps it should be considered that this motivated adversary does have high technical knowledge and analytical means. I do agree, however, that the adversary will not resort to illegal means of re-identification.

Thus, a person without specialized technical knowledge is likely to be discouraged from attempting to re-identify subjects if safeguards are in place. However, many of the cases in which the general public will have access to anonymous databases will be when they are made accessible to the public without restriction. And in these cases, it is necessary to apply higher security measures because not only the inexperienced public will have access, but also more experienced adversaries.

On the other hand, a person with a high level of technical knowledge or expertise will not be discouraged from carrying out a reidentification attack if he knows that he has the capacity to do so. Moreover, certain categories of data are likely to be more attractive than others, either because of their economic value or because of their ability to bring to light private data that can be used to ridicule a person, challenge his ideas, etc. (e.g., search engines, etc.). (for example, the searches that a person performs in his or her web browser). Thus, in my opinion, it would be necessary to consider that the adversary does have relatively high knowledge or means. A deeper analysis would be necessary to determine how to define then the concept and the level of hacking skills and knowledge that should be assumed by our motivated adversary.

Indeed, we find ourselves in a society that is increasingly moving towards transparency - the ultimate expression of which is open data practices - and towards the creation and processing of data - thanks to the Internet of Things and cloud computing. In this scenario, establishing a high-security threshold is a basic requirement.

In any case, the anonymization code of practice developed by the ICO contains very useful recommendations. In the absence of a regulation or an anonymization code developed by EU authorities, some Spanish companies (such as Telefónica[150]) have used this code developed by the ICO to help them develop their internal data anonymization protocols.

[150] Hernandez, Emilcy & Duque, Néstor and Cadavid, Julián. “Big Data: an exploration of research, technologies and application cases”. 20. 17-24. (December, 2017)

In any case, it should be borne in mind that when an organization creates personal data through a re-identification process without the knowledge or consent of the individual, it is obtaining personal data illegally, and as such is liable to be fined.

3.6 Other anonymization techniques

Faced with the challenges of re-identification, complementary techniques have been developed to further limit the possibilities of inferring the identity of the subjects of a dataset, such as differential privacy or k-anonymization.

The techniques encompassed in what is called differential privacy are an important area of research. They are based on the assumption that the risk to a person's privacy should not be increased by the fact that his or her data are stored in a database. That is, it should not be possible to obtain information about an individual from a database that cannot otherwise be known, without access to the database. Thus, for the individual, being present in the databases would not pose an added risk to his or her privacy. This is undoubtedly an area of development that is already being researched, and where the work of Cynthia Dwork[151] stands out.

Anthony Tockar, of Northwestern University, points out that the solution lies in adding noise in related databases so that the identity of each individual is masked while maintaining the usefulness of the data and high accuracy of the information.

Another technique being developed is called k-anonymization. Intuitively, k-anonymization consists of generalizing the attributes of various subjects so that a number of "k" subjects share the same value. Thus, the larger the "k" value, the greater the guarantee of privacy. This prevents a subject from being individualized, at least within that group[152].

[151] Cynthia Dwork. "Differential Privacy". *Microsoft Research.* http://www.audentia-gestion.fr/MICROSOFT/dwork.pdf

[152] Lantaya Sweeney. "K-Anonimity, a model for protecting privacy". *International Journal on Uncertainty*, Fuzziness and Knowledge-Based Systems, Vol. 10, n.o 5 (2002).

Conclusion

After the entry into force of the European Directive, it soon became clear that each of the member countries would transpose it at different rates and with different levels of protection. This, coupled with the fact that technological advances were continuing and that the digital era was expanding, led the Member States to start working on a much more ambitious project: a European Regulation that would be directly applicable in all of them. It's been 4 years since the GDPR has entered into force. And now it is time to analyze and see how successful GDPR has been and what benefits it brings in terms of protecting personal data.

Based on the research done, the European Regulation brings, among others, the following advantages:

- harmonization of legislations.

The main difference between the Regulation and the European Directive is that the former is of direct application in the member countries, not being necessary the transposition of the same, as in the case of a Directive. In this way, irregular and/or fragmented application of European regulations is avoided. Thus, harmonization of legislation is achieved and a more solid and coherent framework for data protection is ensured throughout the European Union, which entails a reinforcement of legal certainty and transparency, generating confidence among citizens and companies[153].

- level of adequacy of protection

The Regulation makes it possible to guarantee a uniform and high level of protection for people in all the countries of the European Union. There is no longer any room for differences in the level of protection between countries, which can be confident that the personal data of their citizens will be treated in the same way in their own country as in another EU country.

[153] European Commission. *"Two years of the GDPR: Questions and answers"* Brussels. (24 June, 2020) https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_1166

This same level of protection means that all Member States provide individuals with the same level of rights and obligations and responsibilities for controllers and processors and equivalent sanctions in all Member States, as well as effective cooperation between the supervisory authorities of the different Member States[154].

- guaranteeing the free movement of data

The GDPR recognizes in its explanatory memorandum that it aims to contribute to the full realization of an area of freedom, security, and justice and to economic and social progress, as well as to the well-being of people [155].

Integration between the countries of the Union has led to a substantial increase in cross-border flows of personal data. Throughout the Union, the exchange of personal data between associations and companies, public and private institutions, including natural persons, has increased. It is therefore necessary to establish a robust, consistent and harmonized regime allowing both private companies and public authorities to use personal data on a large scale, further facilitating the free flow of personal data within the Union and the transfer to third countries and international organizations, while ensuring a high level of protection of personal data.

- guaranteeing the rights of individuals

As we have already seen, the Regulation imposes the same level of protection of the rights and freedoms of natural persons within the Union, so that the processing of such data must be equivalent in all Member States[156]. It ensures throughout the Union that the application of the rules on the protection of the fundamental rights and freedoms of natural persons in relation to the processing of personal data is consistent and homogeneous. The GDPR also recognizes a certain margin of maneuver for the Member States, which they may or may not use. This leeway is provided for in very specific and limited situations.

---

[154] Idem.
[155] See Article 1(3) GDPR.
[156] European Commission. *"Two years of the GDPR: Questions and answers"* Brussels. (24 June, 2020) https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_1166

While talking on the advantages brought forth by GDPR's entry into force, disadvantages coming with it cannot be overlooked.

- purpose limitation

According to Article 5(1)(b) of GDPR, individual data can only be collected for "specific, explicit, and legitimate" purposes and limits the possibility of use of said data for further processes that might arise in the future. This whole point is clearly in contrast with the idea of what Big Data stands for. In most of the situations, data can be used in ways that neither the person/entity who has collected the data nor the person who has shared their data could have even imagined in the beginning. So, in a way, complying with the conditions set in this article limits the ways in which data can lead to so much more than what it was initially intended[157].

- special categories of data

A foundation of EU information assurance strategy is the making of a layered system, where a few types of information classes and datasets are dealt with uniquely in contrast to other people. In the DPD, article 8(1) restricted the handling of information "revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and the processing of data concerning health or sex life," while giving limited exceptions[158]. This differentiation was embraced by the GDPR. The GDPR's article 9 forbids the handling of comparable "special categories," while adding hereditary information, biometric information with the end goal of remarkably distinguishing a characteristic individual, and information connected with sexual direction to the rundown of unique categories[159].

The ascent of Big Data significantly subverts the rationale and utility of applying a different and sweeping lawful system to "special categories" for different reasons. First, there are reasonable considerations. Taking care of this differentiation produces broad and

[157] Zarsky, Tal Z. "Incompatible: The GDPR in the Age of Big Data." *Seton Hall Law Review 47* (2017): 995.
[158] DPD, *supra* note 4, art. 8(1)
[159] GDPR, *supra* note 1, art. 4(13–15) (providing elaborate definitions to the terms "genetic data," "biometric data," and "data concerning health," respectively).

superfluous administrative expenses. Controllers on both the mainland and homegrown level will be expected to consider over the inquiry concerning whether different datasets and examinations fall inside the exceptional classes noted. Courts will likewise have to say something regarding this irrelevant inquiry which will be exorbitant to every one of the gatherings in question. The current regulations will produce significant vulnerability, which will again block on firms trying to participate in Big Data analysis, with companies that have modest budgets experiencing the biggest damages since they cant afford the legal advice. At last, given the symbolic justification to keep up with the differentiation among sensitive and other types of data, it is critical to underline other symbolic motivations to forsake the utilization of exceptional classifications[160].

At a business level, the GDPR has implications for all departments or areas. That is why most companies choose to hire or appoint a data protection delegate to apply additional procedures and guarantees. In addition, it must be someone with the appropriate training to be able to carry out audits and prevent possible sanctions, which may be up to 4% of the annual turnover or 20 million euros (whichever is greater)[161].

With the arrival of the GDPR, data protection has greatly expanded as it gives more rights to people. Rights that must be communicated to the interested parties. Another novelty that is incorporated is the issue of consent. Although EU law has always required that individuals' consent to the collection of their data be free, specific, and informed, the GDPR requires that it be confirmed by a statement or other clear affirmative action[162]. That is, the boxes already checked on the web pages, the silence or inaction of the interested party after reading a privacy statement no longer constitutes their consent.

Another point to note is that people now have the right to move, copy or transfer their personal data from one place to another, including to a competitor. The extension of the scope of application is a point to also take into account. The GDPR makes responsible for security breaches of personal data not only the company that collects it, but also any third party that

[160] Zarsky, Tal Z. "Incompatible: The GDPR in the Age of Big Data." *Seton Hall Law Review 47* (2017): 995.

[161] See Article 83(5) GDPR. "The fine framework can be up to 20 million euros, or in the case of an undertaking, up to 4 % of their total global turnover of the preceding fiscal year, whichever is higher."

[162] See Article 4(11) GDPR. "Consent of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her."

processes it on its behalf, be it another company, an organization, or a natural person[163]. Moreover, it is not enough to simply comply with the GDPR. Companies must demonstrate that they do so in accordance with the "proactive responsibility" requirement, which involves meeting some fairly costly record-keeping obligations.

In addition, throughout the useful life of the personal data, technical and organizational measures must be taken in accordance with the privacy expectations of the data subject (that is, there will be end-to-end privacy of that data). Another novelty is that, in the event of data breaches, the companies that collect personal data must inform the control authorities within 72 hours of their knowledge[164]. It may be also necessary to appoint a data protection officer[165] with knowledge of data protection legislation, although it is not necessary for him to be a direct employee, he can perform this function within the framework of a service contract.

Since the European General Data Protection Regulation is accompanied in particular by detailed transparency, information, and documentation obligations for companies, many companies must identify, analyze and, if necessary, change all processes in the company that involve personal data in order to implement and comply with the GDPR[166].

This goes hand in hand with corresponding personnel and financial effort, which differs from company to company. Companies for which the implementation of and compliance with the GDPR is associated with a higher effort, as well as companies that have fewer resources available to deal with this effort, may experience a competitive disadvantage compared to companies for which this is not the case. Companies that are subject to the GDPR could also be at a competitive disadvantage compared to companies that are not subject to the GDPR and therefore do not have to make this effort[167]. In the case of economic activities within the EU or if data of persons residing in the EU are processed, all companies must comply with the GDPR regardless of their registered office. For activities outside the EU, only companies established

[163] See Article 20 GDPR. "Right to Data Portability".

[164] See Article 33(1) GDPR. "In the case of a personal data breach, the controller shall without undue delay and, where feasible, not later than 72 hours after having become aware of it, notify the personal data breach to the supervisory authority competent in accordance with Article 55, unless the personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons. Where the notification to the supervisory authority is not made within 72 hours, it shall be accompanied by reasons for the delay.

[165] See Article 37(1) GDPR. "The controller and the processor shall designate a data protection officer in any case……"

[166] Information commissioner's office (ICO, UK data protection authority). "Guide to the General Data Protection Regulation (GDPR)" (2 August, 2018).

[167] Voss, W. & Houser, Kimberly. "Personal Data and the GDPR: Providing a Competitive Advantage for U.S. Companies". American Business Law Journal. 56. 287-344. (May 2019)

in the EU have to comply with the GDPR. However, the GDPR could also lead to competitive advantages: Compliance could increase trust among customers and cooperation partners, provided they value a high level of data protection[168]. In addition, dealing with the data processes in the company within the scope of the implementation of the GDPR could lead to an improvement of these processes.

Both competitive advantages and competitive disadvantages are conceivable from a theoretical point of view. Whether the GDPR leads to more legal certainty because it harmonizes different regulations or to more legal uncertainty because companies may not understand it is also not clear.

In the same way, customers are more secure regarding the treatment of their information, which has allowed their degree of trust in companies to increase. Those that have correctly followed the legal framework have also benefited from the regulation, being able to demonstrate their transparency and involvement with users' private data, which has resulted in an improvement in their reputation. Research carried out by Cologne Institute for Economic Research (Institut der Deutschen Wirtschaft)[169] also supports this point. The most frequently cited benefit is that of gaining (potential) customers through a high level of data protection (79 percent; Table 1)[170]. One possible explanation for this observation could be that customers attach great importance to data protection and, against this background, consciously choose companies (Selligent, 2019)[171] that focus on this interest and even advertise that they attach particular importance to data protection - possibly going beyond the provisions of the GDPR. 60 percent of companies also see data protection as a selling point to cooperation partners. 40 percent of companies see direct competitive advantages over competitors within or outside the EU. Only 26 percent see the advantage of better use of data, for example by being forced to deal with data-driven processes and the resulting opportunity to improve them.

---

[168] Laszlo Delle. “GDPR Compliance as a Competitive Advantage” ISACA Blog, (16 January 2019).

[169] Barbara Engels, Marc Scheufen “Eine Analyse basierend auf einer Befragung unter deutschen Unternehmen: Wettbewerbseffekte der Europäischen Datenschutzgrundverordnung” *IW-Report 1/20.* (15 January, 2020). [self-translated]

[170] Idem.

[171] Selligent. “The Customer’s Perception Is Your Reality”. Selligent Global Connected Consumer Index, 3rd Edition. (2019). https://www.selligent.com/wp-content/uploads/2021/10/white-paper-connected-consumer-index-2020-us.pdf

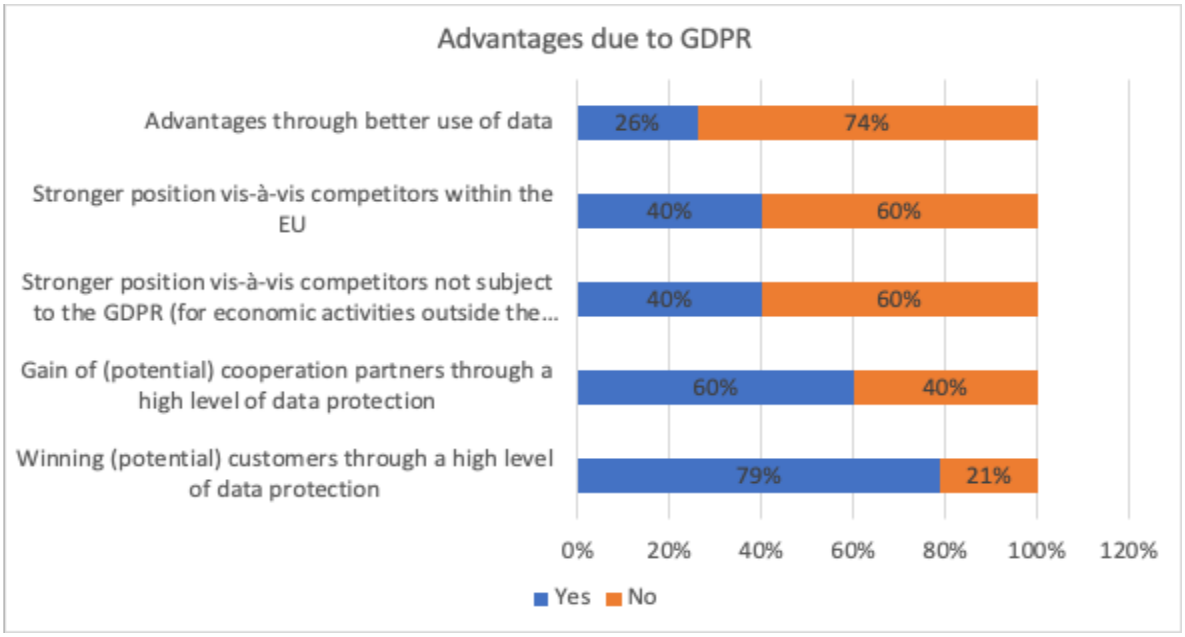


Table 1. Proportion of companies that see the respective aspect as an advantage of the GDPR out of all companies that consider the GDPR to be beneficial for competition, in percent

Moreover, the obligation to review databases has been a great opportunity for many organizations that needed to digitize their information. A large number of companies decided to take advantage of the occasion by working on GDPR compliance to move all information to the cloud[172].

But not all have been advantages, since the application of the GDPR has also meant that organizations have had to face a large workload, from sending thousands of emails to obtain user consent to reviewing processes to certify correct data collection. Millions of people forgot or ignored the emails to give their consent to the registration of their information, causing the databases of all organizations to lose many contacts who were really interested in being registered in them. Among the disadvantages, the high cost of implementation very clearly dominates. Ninety-six percent of the companies that perceive the GDPR to be a disadvantage for competition cite effort as a disadvantage (Table 2)[173]. Particularly worthy of mention is the disadvantage of legal uncertainty, which was cited by 89 percent of the companies.

The harmonization of the different data protection levels in the European Union through the GDPR should actually eliminate legal uncertainty. However, it is conceivable that this will

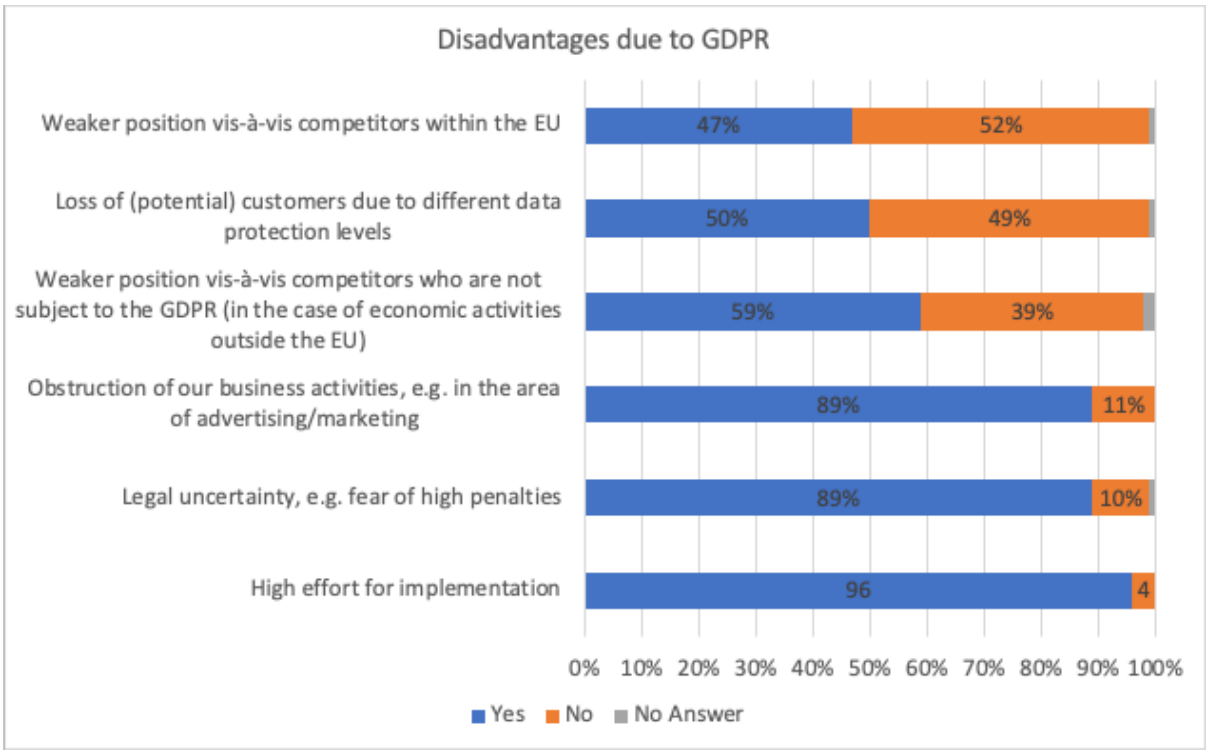
[172] Capgemini. "Championing Data Protection and Privacy – a source of competitive advantage in the digital century". *Capgemini Research Institute*. (September, 2019). https://www.capgemini.com/de-de/wp-content/uploads/sites/5/2019/09/Report_GDPR_Championing_DataProtection_and_Privacy.pdf

[173] Barbara Engels, Marc Scheufen "Eine Analyse basierend auf einer Befragung unter deutschen Unternehmen: Wettbewerbseffekte der Europäischen Datenschutzgrundverordnung" *IW-Report 1/20.* (15 January, 2020). [self-translated]

only become apparent in the medium to long term, when companies have become accustomed to the new regulation and court rulings and proceedings based on the GDPR supplement the theory of the GDPR with practice and thus make the regulation more precise.

In addition, 89 percent of companies even see the GDPR as an impediment to their own business activities[174]. This is likely to be the case in particular for those companies that use personal data as the basis for new and innovative data-driven business models. However, companies whose marketing measures are heavily based on personal data - and this is likely to be the case frequently in times of personalized online advertising - are also affected by this disadvantage, because the GDPR sets strict limits on the exploitation of personal data and requires consent for data processing.

In addition, although at the European level the GDPR has meant a regulatory unification to facilitate the work of multinationals, in practice, it has meant a new bureaucratic procedure for both users and companies. Many services such as online video platforms or financial products need user data to operate, which is why the GDPR has forced many companies to redesign and adapt both registration systems and the products themselves[175].



[174] Idem.

[175] Larsson, Anthony & Lilja, Pernilla. “GDPR: What are the risks and who benefits?”. (September, 2019) https://www.researchgate.net/publication/337305085_GDPR_What_are_the_risks_and_who_benefits

Table 2. Percentage of companies that see the respective aspect as a disadvantage of the GDPR out of all companies that see the GDPR as disadvantageous for competition

Having seen the incipient but growing synergy that is currently taking place between private powers and the States when using the personal data that users generate on the network, the conclusions that we can draw are from different fields. In general, we can affirm that the digitization of information confronts us with a series of challenges that end up being unavoidable in modern capitalist societies and undoubtedly affect the power balances of the world. Schneier's hypothesis is that, given the increase in power derived from technological innovation, at a time of rapid change, this power is often used by social actors to act for their own benefit, creating new social dilemmas that require action. equally fast and efficient if balances of power are to be guaranteed[176]. At a time when technology changes so rapidly, citizen privacy protection mechanisms should also adapt quickly, but to do so in time, future laws would have to be valid for both real and international environments. by digital environments, since otherwise they would have to be continuously reformulated by the emergence of new risks that technology facilitates. In this context, the mechanisms for its defense must be rethought, adapting to the new circumstances that these technologies have made emerge. Thus, as technical conclusions, we would refer to the need to understand the functioning of the mechanisms that operate in all this phenomenon. Whether to protect the right to privacy of the users or to increase their trust in the network, a simplification of privacy policies is necessary, to facilitate that they make conscious and informed decisions regarding the disclosure or exposure of their information. We find that, although most users and consumers use the Internet in more and more aspects of their lives, many have a lack of privacy policies, and it is also foreseeable that this fact is favored by the appearance of new technological devices. Hence, the GDPR's entry into force has been, in many ways, game-changing. Giving the consumers more control over their data, how much of it they can share and the possibility of withdrawing consent at any time has increased the trust. This, in turn, has benefited the companies that rely on data to avoid complications over data privacy regulations. The harmonization of the rules and having a single point of reference when it comes to the data regulations for EU has greatly increased customers disclosing their personal data. GDPR does

[176] Bruce Schneier. "Liars and Outliers: Enabling the Trust that Society Needs to Thrive and Carry On" *Indianapolis: John Wiley & Sons* (2012).

come with its own problems that it creates for companies as in increased costs and more layers added to the bureaucracy. Moreover, problems that arise from GDPR rules such as "purpose limitation" prevent Big Data to do what it does best: get creative and produce solutions[177].

[177] Kapoor, Hansika & Tagat, Anirudh. "Everything counts: Big Data and creativity science". 70 Years of Research into Creativity: JP Guildford's Role and Today's Focus (pp.125-143). 2020.

Bibliography

BOOKS AND ARTICLES

1. Aluja, Tomas. “Data mining, between statistics and Artificial Intelligence”. *Polytechnic University of Catalonia, Vol. 25, n.o 3.* 2001.
2. Anderson, Richard. “How Mathematicians Dominate the Markets”. *BBC Economy.* October 2011.
3. Baek, Young Min & Bae, Young & Jeong, Irkwon & Kim, Eunmee & Rhee, June. “Changing the default setting for information privacy protection: What and whose personal information can be better protected?”. *The Social Science Journal. 51.* July 2014.
4. Barocas, Solon and Nissembaum, Helen. “Privacy, big data and the public good»; Chapter 2: Big data's End Run Around Anonymity and Consent”, *Cambridge University Press.* 2014.
5. Batta, Mahesh. “Machine Learning Algorithms - A Review”. *International Journal of Science and Research (IJSR)*. January 2019
6. Bellman, Steven & Johnson, Eric & Lohse, Gerald. “To Opt-in or Opt-out? It Depends on the Question”. *Commun. ACM. 44. 25-27*. February 2001
7. Bottles, K., Begoli, E., & Worley, B. “Understanding the Pros and Cons of Big Data Analytics”. *Physician Executive, 40(4), 6-12.* July 2014.
8. Boyd, Danah and Crawford, Kate. “Critical questions for Big Data” *Information, Communication and Society. 662-679*. 2012.
9. Boyd, Danah. “Networked privacy”. *Personal Democracy Forum, New York.* 2011.
10. Capgemini. “Championing Data Protection and Privacy – a source of competitive advantage in the digital century”. *Capgemini Research Institute.* September 2019. https://www.capgemini.com/de-de/wp-content/uploads/sites/5/2019/09/Report_GDPR_Championing_DataProtection_and_Privacy.pdf
11. Castells, Manuel. “The Information Age: Economy, Society and Culture: The Network Society”. 1999.

12. Cate, Fred H. “The failure of Fair Information Practice Principles”. *Consumer Protection In The Age Of Information Economy.* 2006.
13. Cate, Fred H., and Mayer-Schönberger, Viktor. “Notice and consent in a world of Big data”. *International Data Privacy Law, Vol 3, n.o 2.* 2013.
14. Clarke, Roger. “Big data, big risks”. *Information Systems Journal. 26. 77-90.* January 2016.
15. OECD, Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data. (1981) https://rm.coe.int/1680078b37
16. Cukier, Kenneth and Mayer-Schonberger, Viktor. “The Rise of Big data. How It's Changing the Way We Think About the World”. *Foreign affairs Vol. 92, No. 3.* 2013.
17. Cumbley, Richard and Church, Peter. “Is “Big Data” Creepy? *Computer Law & Security Review, 29, 601-609.* 2013.
18. Dalessandro, Brian. “The science of privacy” Ad: *Tech (blog).* July 2013.
19. Delgado Rodríguez, Miguel and Llorca Díaz, Javier. “Longitudinal studies: concept and characteristics”. *Spanish magazine of Public Health, Vol. 78, no. 2.* 2004.
20. Dellei, Laszlo. “GDPR Compliance as a Competitive Advantage”. *ISACA Blog*. January 2019.
21. Domingo-Ferrer, Josep. “Big Data Anonymization Requirements vs Privacy Models”. *471-478.* (2018).
22. Domingo-Salvany, Antonia and Brugal Puig, Maria Teresa and Barrio Anta, Gregorio. “SET Treaty of Addictive Disorders”. *Spanish Society of Addictions. Editorial Médica Panamericana.* 2006.
23. Dwork, Cynthia. “Differential Privacy”. *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006).* July 2006
24. El Amam, Khaled and Alvarez, Cecilia. “A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques”. *International Data Privacy Law Journal, Oxford University Press, Vol. 5, n.o 1.* 2015.
25. Engels, Barbara and Scheufen, Marc “Eine Analyse basierend auf einer Befragung unter deutschen Unternehmen: Wettbewerbseffekte der Europäischen Datenschutzgrundverordnung” *IW-Report 1/20*. 15 January, 2020. [self-translated]

26. Finger, Lutz. “Recommendation Engines: The Reason Why We Love Big data”. *Forbes Tech.* September 2014.
27. Fish, Isabella. “Drivers' phone data is being used by York City Council to try and ease the city's traffic jams and could be extended across the country”. *Daily Mail.* April, 2018.
28. Gainsbury, Sally & Angus, Douglas & Procter, Lindsey & Blaszczynski, Alex. “Use of Consumer Protection Tools on Internet Gambling Sites: Customer Perceptions, Motivators, and Barriers to Use”. *Journal of Gambling Studies.* March 2020.
29. Galli, Ricardo. “Be careful with Big Data” *Free Software Blog internet, legal.* May 2013.
30. Goes, Paulo. “Big Data and IS Research”. *MIS Quarterly, 38(3), iii-viii.* 2014.
31. Hardy, Quentin. “Rethinking privacy in an era of big data”. *New York Times.* June 4, 2012.
32. Hernandez, Emilcy & Duque, Néstor and Cadavid, Julián. “Big Data: an exploration of research, technologies and application cases”. *TecnoLógicas 20. 17-24.* December, 2017.
33. Horvát, Emöke-Ágnes & Hanselmann, Michael & Hamprecht, Fred A. & Zweig, Katharina A. “One plus one makes three (for Social Networks)”. *PLOS One Journal.* 2012.
34. “How Big Data Analysis helped increase Walmarts Sales turnover?”. 25 January 2022. https://www.projectpro.io/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109
35. IBM Institute for Business Value, in collaboration with the Said School of Business at the University of Oxford. “Analytics: Using Big Data in the Real World”. *IBM Global Business Services.* 2012.
36. James, Konow. “Coercion and Consent”. *Journal of Institutional and Theoretical Economics JITE*. March 2014.
37. Jernigan, Carter and Mistree, Behran F. T. “Gaydar: Facebook friendships expose sexual orientation”. *First Monday, Vol. 14, n.o 10.* 2009.

38. Kapoor, Hansika & Tagat, Anirudh. “Everything counts: Big Data and creativity science”. *70 Years of Research into Creativity: JP Guildford’s Role and Today’s Focus (pp.125-143).* 2020.
39. Kuner, Christopher & Cate, Fred & Millard, Christopher & Svantesson, Dan. “The challenge of 'big data' for data protection”. *International Data Privacy Law. 2. 47-49.* 2012.
40. Larsson, Anthony & Lilja, Pernilla. “GDPR: What are the risks and who benefits?”. *Routledge.* September 2019.
41. Lohr, Steve. “If Algorithms Know All, How Much Should Humans Help?” *The New York Times News Services*. April 2015.
42. Mayer-Schonberger, Viktor and Cukier, Kenneth. “The big data revolution”. 2013.
43. McDonald, Aleecia and Cranor, Lorrie Faith. “The Cost of Reading Privacy Policies. Journal of Law and Policy of the Information Society”, Vol. 4, n.o 3. 2008.
44. Misolve, Alan et al. “You are who you know: inferring users profiles in online social networks”. *Web Search and Data Mining (WSDM). ACM, New York*. 2010.
45. Montuschi, Luisa. “Troubled ethical issues in the information age, internet and the world wide web”. *CEMA Working Papers, University of CEMA.* 2005.
46. Mooij, Joris M. & Peters, Jonas & Janzing, Dominik & Zscheischler, Jakob & Scholkopf, Bernhard. “Distinguishing cause from effect using observational data: methods and benchmarks”, version 2. *Journal of Machine Learning Research, Cornell University.* 2015..
47. Narayanan, Arvind and Shmatikov, Vitaly. “De-anonymizing social networks”. *The University of Texas at Austin, Simposio IEEE on Security and Privacy.* 2009.
48. Narayanan, Arvind and Shmatikov, Vitaly. “Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)”. *The University of Texas at Austin.* 2008.
49. Nunan Dan and Yenicioglu Baskin. “Informed, Uninformed and Participative Consent in Social Media Research”. *International Journal of Market Research. 55. 791.* January 2014.
50. O'Connor, Fred. “Google Flu Trends calls out sick, indefinitely”. *PCWorld.* August 2015.

51. Ohm, Paul. “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization”. *UCLA Law Review, Vol. 57.* 2010.
52. Puyol, Javier. “An approach to big data”. *Journal of Law of the National University of Distance Education (UNED), No. 14.* 2014.
53. Rubinstein, Ira S. “Big data: The End Of Privacy Or A New Beginning?” *International Privacy Law, Vol. 3, n.o 2.* 2013.
54. Schneier, Bruce. “Liars and Outliers: Enabling the Trust that Society Needs to Thrive and Carry On” *Indianapolis: John Wiley & Sons.* 2012.
55. Selligent. “The Customer’s Perception Is Your Reality”. *Selligent Global Connected Consumer Index, 3rd Edition.* 2019. https://www.selligent.com/wp-content/uploads/2021/10/white-paper-connected-consumer-index-2020-us.pdf
56. Sweeney, Lantaya. “K-Anonymity, a model for protecting privacy”. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10, n.o 5.* 2002.
57. Szpunar, Maciej. “Reconciling new technologies with existing EU law – Online platforms as information society service providers”. *Maastricht Journal of European and Comparative Law. 27. 399-405.* August 2020.
58. Tene, Omer and Polonestsky, Jules. “Big data for all: Privacy and user control in the age of analytics”. *Northwestern Journal of Technology and Intellectual Property, Vol. 11, n.o 5.* 2013.
59. Tene, Omer and Polonestsky, Jules. “Privacy in the age of big data: a time for big decisions”. *Standford Law Review Online, Vol. 64, n.o 63.* 2012.
60. Toffler, Alvin. “The Third Wave” p179. *New York: William Morrow and Company, Inc.* 1980.
61. Urrutia, Ana Victoria Sanchez, & Gorski, Héctor Claudio Silveira, & Michel, Mónica Navarro, & Rodota, Stefano. “Technology, privacy and democratic society”. *Icaria.* 2003.
62. Voss, W. & Houser, Kimberly. “Personal Data and the GDPR: Providing a Competitive Advantage for U.S. Companies”. *American Business Law Journal. 56. 287-344.* May 2019.

63. Wang, Lidong and Alexander, Cheryl. “Big Data Analytics in Healthcare Systems”. *International Journal of Mathematical, Engineering and Management Sciences. 4. 17-26.* January 2019.
64. World Economic Forum and the Boston Consulting Group. “Rethinking Personal Data: Strengthening Trust”. 2012.
65. Zarsky, Tal Z. “Incompatible: The GDPR in the Age of Big Data.” *Seton Hall Law Review 47* (2017): 995.

LEGISLATION AND OPINIONS OF DATA PROTECTION BODIES

1. Article 29 Working Party. "Opinion 03/2013 on Purpose Limitation" (2013). https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf
2. Article 29 Working Party. “Opinion 05/2014 on Anonymisation techniques” (2014). https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf
3. Article 29 Working Party. 15/2011 on the definition of consent (WP 187) pp.19-20. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf
4. Article 29 Working Party. Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (WP251), paragraph IV.B, p. 20 onwards. https://ec.europa.eu/newsroom/article29/items/612053
5. 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 *OJ L 281, 23.11.1995, p. 31–50*
6. Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications). *OJ L 201, 31.7.2002, p. 37–47*
7. European Commission. “Two years of the GDPR: Questions and answers” Brussels. 24 June, 2020. https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_1166
8. European Commission Staff Working Paper, Impact Assessment, Annex 2, p. 20 and also pp. 105-106.

9. European Union European Commission, European Digital Single Market. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52015DC0192&from=EN
10. European Union, 2000 Charter of Fundamental Rights of the EU. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012P/TXT&from=EN
11. European Union, Europe 2020 Strategy of the European Council. https://ec.europa.eu/eu2020/pdf/COMPLET%20EN%20BARROSO%20%20%20007%20-%20Europe%202020%20-%20EN%20version.pdf
12. Federal Trade Commission (FTC). "Protecting Consumer Privacy in an Era of Rapid Change. Recommendations for Businesses and Policymakers". 2012.
13. Health Insurance Portability and Accountability Act Privacy Rule (HIPAA), the United States. 1996.
14. Information Commissioner's Office (ICO, UK data protection authority). "Big Data and Data Protection". 2014.
15. Information commissioner's office (ICO, UK data protection authority). "Anonymisation: managing data protection risk code of practice". 2012.
16. Information commissioner's office (ICO, UK data protection authority). "Guide to the General Data Protection Regulation (GDPR)". August 2018.
17. Opinion of the European Data Protection Supervisor of 14 January 2011 on the Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - "A comprehensive approach to the protection of personal data in the European Union."
18. Personal Health Information Protection Act, Ontario, Canada (2004).
19. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *OJL 119, 4.5.2016, p. 1-88.*
20. Treaty of Lisbon amending the Treaty on European Union and the Treaty establishing the European Community, signed at Lisbon, 13 December 2007 OJ C 306, 17.12.2007, p. 1–271.

AUDIOVISUAL RESOURCES AND WEBSITES

1. Marr, Bernard. “What Is Unstructured Data And Why Is It So Important To Businesses? An Easy Explanation For Anyone” [Website]. Bernard Marr & Co (2013) https://bernardmarr.com/what-is-unstructured-data-and-why-is-it-so-important-to-businesses-an-easy-explanation-for-anyone/
2. Puyol, Javier. “Big data and Public Administrations” [Conference]. International Seminar Big data for Official Information and Decision Making (2014).
3. Elliot, Mark. “To be or not to be (anonymous)? Anonymity in the age of big and open data”. [Conference] Computers, Privacy & Data Protection on the Move (CPDP) (2015).
4. Panopticlick. Project “How unique —and trackable— is your browser?” Available in: https://coveryourtracks.eff.org/
5. SeedScientific. website: https://seedscientific.com/how-much-data-is-created-every-day/#:~:text=How%20much%20content%20is%20created,2.5%20quintillion%20bytes%20of%20data.