Chapter 9: Beyond the Access-Use Debate: Regulating the Analysis of Information in the Big Data Era in Order to Ensure Reliability and Trustworthiness

Bart van der Sloot

Abstract

Companies and governmental organizations increasingly rely on Big Data technologies. They anticipate that this will increase the effectiveness of operations, reduce costs and optimize the reliability of procedures. Although it is impossible to give a precise definition, Big Data processes typically run through three phases: gathering data, analysing data and using data. The current legal regime focusses primarily on the first phase, setting limits on which data organizations are allowed to gather, how much data can be gathered and for what reasons. It has been suggested to let go of this focus and instead introduce regulation on the use of data. This would remove the hurdles for data-driven innovation and at the same time focus on preventing harms following from Big Data usage, which they feel is currently under-regulated. Others fear that use-based regulation would be inadequate in preventing harms and prefer the current access-based regulation, among others because without barriers to gathering data, data-exploitation would explode and because proving that harm derives directly from data usage is often problematic. Moving beyond this so-called access-use debate, this chapter suggests to look at the second phase of Big Data processes, during which data are stored, categorized and analysed. It is this phase where most errors occur and at the same time, where no or very limited rules and regulations exist. Introducing standards for analysing data would optimize the analysis of data and hence increase the reliability of data analytics and the trust of citizens in governmental agencies relying on such data analytics.

9.1. Introduction

The reliability and trustworthiness of governmental organizations is not only important to ensure legitimacy, but also their effectiveness. People that feel that the government is unreliable, unpredictable and not trustworthy will typically have a more relaxed approach towards the law and will be less willing to cooperate with official institutions. This holds true, *inter alia*, for tax authorities, as has been pointed out by Benno Torgler (2003), stressing that

> if taxpayers trust the public officials, they are more willing to be honest. If the government acts trustworthily, taxpayers might be more willing to comply with the taxes. The relationship between taxpayers and government, similar to the one with the tax administration can be seen as a relational contract or psychological contract, which involves strong emotional ties and loyalties.

Currently, many governmental organizations are experimenting with Big Data projects, in which large quantities of data are gathered, analysed through smart algorithms and used for policy design and decision-making. They anticipate that this will ensure a more effective use of governmental power, time and manpower and hence reduce costs. So far, however, Big Data has yielded limited success. Experts argue that relying on mass surveillance for preventing terrorism is simply not the right tool (Schneier, 2016), predictive policing projects around the world have been stopped due to a lack of results (Saunders, Hunt & Hollywood, 2016), and attempts of tax authorities to make a shift towards the data-driven future have not been overwhelmingly successful (Algemene Rekenkamer, 2017). This means that in many instances, Big Data projects have been cancelled after a number of months or years and that data-driven applications have in fact cost more money and manpower, instead of reducing the costs. Time will tell whether Big Data will once be as effective as the gurus predict.

There is a different problem involved with Big Data, which will be the core of this chapter, namely the reliability of data analytics and hence the reliability and trustworthiness of governmental agencies that rely on such analytics (*see* further: Boyd & Crawford, 2011, 2012). Big Data can be divided roughly in three phases: gathering data, analysing data and using data. The current legal regime is particularly focussed on setting rules and limits on the first phase, specifying which data organizations are allowed to gather, how much data can be harvested and for what reasons.

A number of experts have suggested to focus on the last phase of Big Data processes in which the data are used instead. This would remove the hurdles for data-driven innovation currently in place and at the same time focus on preventing harms following from Big Data usage. This phase is currently under-regulated while the data-driven harms, according to these experts, is the real problem of Big Data applications. Defenders of the current access-based regulation, to the contrary, fear that use-based regulation would be inadequate in preventing harms and remove the dams in place, creating a data flood.

Moving beyond this so-called access-use debate, this chapter suggests a look at the second phase of Big Data processes, when data are stored, categorized and analysed. It is this phase where most errors occur and at the same time, where no or very limited rules and regulations exist. Introducing standards for analysing data would optimize the analysis of data and hence increase the reliability of data analytics. This would in its turn have a positive effect on the trustworthiness and reliability of government agencies relying on Big Data usage and the confidence of citizens in their government and the exercise of power by governmental agencies.

The structure of this chapter is simple. Section 9.2 will suggest that Big Data processes can be divided into three phases: gathering data, analysing data and using data. Section 9.3 will show that the current legal framework is in particular focussed on setting rules and limits on the first phase of data processes, in which data are gathered; experts have proposed to replace this focus by a use-based approach to legal regulation, focussing on the third phase of Big Data processes. Section 9.4 will show, however, that in particular in the second phase, the moment in which the data are analysed, there are many problems and dilemma's to be resolved; it is also this phase that is currently left mostly unregulated. Section 9.5 will gather and summarize a number of building blocks which may be used for regulating the second phase of Big Data processes, in which data are analysed and patterns and profiles are made. Section 9.6, finally, will make some initial suggestions on which standards could be implemented to ensure more reliable data analytics.

9.2. Big Data

Big Data as a concept and phenomenon is hard to pinpoint. There are four reasons why attempts to determine whether a new technique can be qualified as Big Data or not will be unsuccessful (*see* more in detail: Wetenschappelijke Raad voor Regeringsbeleid, 2016):

1. *Umbrella Term:* Big Data is often used as a umbrella term for all kinds of new technologies and developments, such as Open Data (the idea that data should be open and accessible for everyone and not privatized by a certain organization or person) (Kitchin, 2014); Re-Use (the belief that data can always be used for new purposes/that data can always have a second life) (van der Sloot, 2011); Algorithms (the computer programs used to analyse the data and produce statistical correlations) (Pasquale, 2015); Profiling (Big Data analysis usually makes use of categories and predictions, such as that 20% of the people with a red car also like rap music, that 40% of the men that have a house with a value over € 250,000 are over the age of 50, or that 0.01% of the people that go to Yemen for vacation visit extreme terrorist websites and are between the age of 15 to 30 are potential terrorists) (Custers, 2004); Internet of Things (IoT) (the trend to install sensors on objects and connect them to the internet, so that the chair, street light and vacuum cleaner can gather data about their environment) (Witkowski, 2017); Smart Applications (the trend of making the internet-connected devices resonate with their environment and letting them make independent choices – the smart street light that shines brighter when it rains, the smart refrigerator that orders a new bottle of milk, the smart washing machine that turns on whenever the power usage in the area is low, etc.) (Hildebrandt, 2015; Shapiro, 2005); Cloud Computing (the fact

that data can be stored globally and can be located in Ghana at one moment in time and in Iceland the next day, when there is storage capacity in that specific data centre at that specific moment) (Qiang, Zheng & Hsu, 2015); Datafication (the trend to rely more on data about reality than on reality itself) (Mayer-Schönberger & Cukier, 2013); Securitization (the usage of risk profiles in order to prevent potential threats from materializing) (Beck, 1992); Commodification (the trend to commercially exploit data) (Crain, 2018); Machine Learning (the possibility that machines and algorithms can learn independently – *e.g.* deep learning, through which algorithms can learn from their environment and adapt beyond how they have been programmed to behave) (Murphy, 2012); and Artificial Intelligence (AI) (the trend to rely more and more on computer intelligence, with some experts saying that AI will outplace human intelligence in a decade or two) (Russell & Norvig, 2010). In the various definitions of official institutions, these and other elements are incorporated in the definition of Big Data, making Big Data an umbrella term for all kinds of new technologies and trends rather than a concept in its own right.

2. *Node for Various Developments:* Big Data is also a term used to bring together various developments, such as most obviously technological evolutions, through which the gathering and storing of data is increasingly easy, through which analysing information can be done increasingly swift and yield ever more valuable results and through which these outcomes can be used for innovative application in more and more fields of life. In addition, there are economic developments underlying the concept of Big Data, such as that the costs for these technologies are declining every year, which has led to a 'democratization' of the data applications; in addition, the costs for gathering, storing and analysing data are so low that the economic rationale is seldom a reason to stop or abstain from gathering and storing data. To provide a final example, Big Data is also used to address several societal tendencies, such as those mentioned above: datafication, commodification, securitization, etc.
3. *Historical Fluid:* There is no specific moment in time when Big Data was invented. Rather, most elements used for current Big Data applications have existed for decades. What has changed is that the data technologies have become quicker, the databases bigger and the reliance on data-analysis firmer, but these are gradual changes. What we call Big Data now will be small data in a decade or so.
4. *Fluid by Definition:* Finally, there is no standard definition of Big Data. The most commonly used definition is the so-called 3V model,[1] in which the phenomenon is defined in terms of Volume (the amount of data), Variety (the number of data sources) and Velocity (the speed

[1] *See*, https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

> at which the data can be analysed). In addition, others have added Vs, such as Value (Dijcks , 2012) (the value Big Data represents), Variability (Hopkins & Evelson , 2011) (the speed at which Big Data processes change), and Veracity[2] (the exactitude of Big Data processes). What is mutual to all these elements is that they are gradual. There is not a precise moment at which the database becomes so big that one can speak of Big Data – *e.g.* that a database of 1,000 data points is not Big Data but a database with 1,001 data points would be. The same counts for the other elements. There is not a specific moment at which the data analysis is so quick that it can be called Big Data or a moment at which there are so many different data sources that one can speak of Big Data, *e.g.* that 10 data sources is not Big Data but 11 would be. The other way around, not all of these elements should necessarily be fulfilled to speak of Big Data. For example, some data sets that are big and are analysed with smart algorithms at rapid speed, but only derive from one data-source are still called Big Data. Rather than defining Big Data, Big Data should be treated as an ideal type – the more data are gathered, the more data sources are merged, the higher the velocity at which the data are analysed, etc., the more a certain phenomenon approaches the ideal type of Big Data.

Although it is consequently impossible to give an exact definition of Big Data, it is possible to discern three phases through which Big Data processes go. These are: gathering the data, analysing the data and using the data. These three stages are not unique to 'Big data', but apply to 'data' management in general – still, in each phase, the contrast between Big Data and handling small data becomes clear.

1. *Gathering*: With regard to the volume of data, the basic philosophy of Big Data is 'the more, the merrier'. The larger the data set, the more interesting patterns and correlations can be found, the more valuable the conclusions drawn therefrom. What is said to set Big Data technologies apart is precisely that, relying on smart computers and self-learning algorithms, they can work on extremely large sets of data. By being confronted with a constant stream of new data, these programs can continue to learn and become 'smarter'. It is important that Big Data can not only be used for the collection of data, but also for the production of data. This is done through inferring new data from old data. With respect to the variety of data sources, it should be underlined that Big Data facilitates linking and merging different data sources. For example, an existing database can be linked to a database of another organization and subsequently enriched with information that is scraped from the internet. Big Data is also said to work well on so-called unstructured data, that are uncategorized. Because Big Data is essentially about analysing very

[2] *See*, www-01.ibm.com/software/data/bigdata/what-is-big-data.html

large amounts of data and detecting general and high-level correlations, the quality of specific data is said to become less and less important – quantity over quality. Because data gathering and storage is so cheap, data are often gathered without a predefined purpose, only determining afterwards whether data represent any value and if so, for what purposes they can be used (*see* further: Bollier, 2010; Craig & Ludloff, 2011; Kerr & Earle, 2013; Madden, 2012; McAfee & Brynjolfsson, 2012; Stevenson & Wagoner, 2014; Toh a& Platt, 2013; Young, 2015).

2. *Analysing:* Once the data have been collected, they will be stored and analysed. The analysis of the data is typically focussed on finding general characteristics, patterns and group profiles (groups of people, objects of phenomena). General characteristics can be derived, for example regarding how earthquakes typically evolve, which indicators can be designated that can predict an upcoming earthquake and which type of building remains relatively undamaged after such disasters. An important part of Big Data is that the computer programs used for analysing data are typically based on statistics – the analysis revolves around finding statistical correlations and not around finding causal relations. The statistical correlations usually involve probabilities. It can thus be predicted that of the houses built with a concrete foundation, 70% will remain intact after an earthquake, while of the houses without a concrete foundation, this only holds true for 35%, or that people that place felt pads under the legs of their chairs and tables on average repay their loan more often than people who do not. This also brings the last point to the fore, namely that with Big Data, information about one aspect of life can be used for predictions about whole other aspects. It may appear that the colour of a person's couch has a predictive value for his health, that the music taste of a person's friends on Facebook says something about his sexual orientation, or that the name of a person's cat has a predictive value for his career path (*see* further: Crawford & Schultz, 2014; Davis & Patterson, 2012; Lazer, Kennedy, King & Vespignani, 2014; Payton & Claypoole, 2014; Richards & King, 2013, 2014; Satija & Hu, 2014; Tene & Polonetsky, 2012, 2013, 2017).
3. *Usage:* The correlations gained from Big Data analysis can be used at the general level, for example when policy choices are based on the prediction that in 20 years' time, halve of the population will be obese; they can be used to make predictions about groups of people, events or objects, such as bridges, immigrants or men with red cars and big houses; and they can be applied to specific, individual cases, projecting the general profile on a specific case (*see* further: Dusseault, 2013; Floridi, 2012; Lyons, 2014; Lukoianova & Rubin, 2014; Puschmann & Burgess, 2014; Rubenstein, 2013).

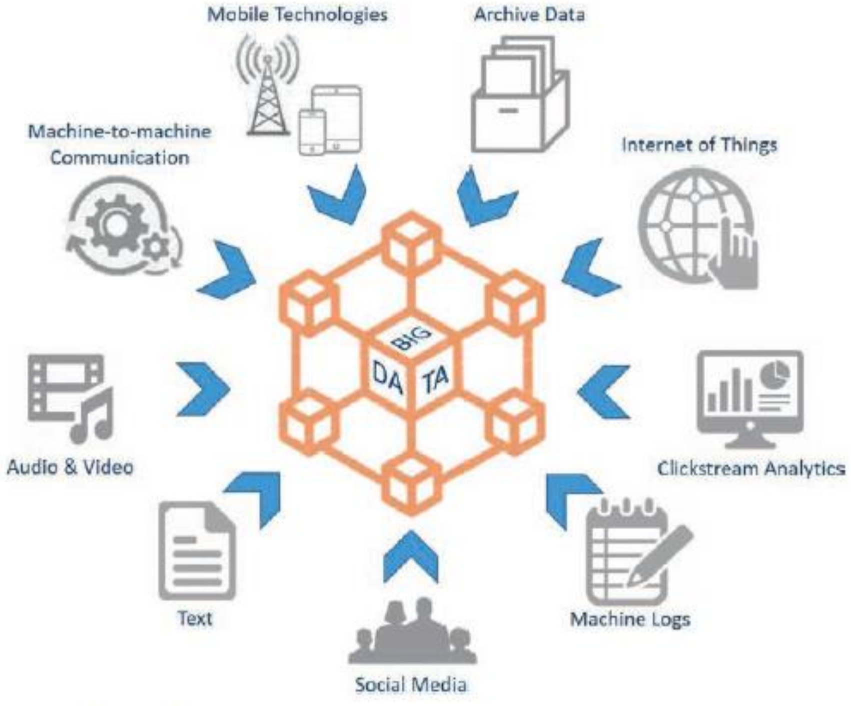
Tax authorities are also increasingly relying on Big Data and such initiatives are promoted, among others, by the Organisation for Economic Co-operation and Development (OECD, 2016a). The envisioned application of such initiatives runs along the lines of what has just been described. Gathering large sets of data, enriching data sets with online information sources, analysing those data to find patterns and profiles, especially to assess the risks involved in certain processes or vis-à-vis certain citizens or companies, and these data findings are applied and used in the workflow and processes of the tax authorities, who increasingly move towards data-driven operations.

Traditionally, the bulk of the data of tax authorities was delivered in a structured form.

> Typically, the bulk of data which is supplied by taxpayers and tax intermediaries arrived and continues to arrive in a structured form specified by tax administration generally in accordance with legislation or regulation. Unstructured data sets were infrequently accessed or supplied, often because revenue bodies lacked the tools to unlock the value from such data or to be able to present it in a form usable by their administration in core tax activities. (OECD, 2016b)

However, more and more tax authorities are experimenting with enriching data sets and gathering data from a variety of sources, such as from audio and video, from the Internet of Things, clickstream data, data gathered from social media and mobile and satellite data (Figure 9.1).

Figure 3.1. Sources of Big Data



Source: Russian Federal Tax Service.

Figure 9.1: Sources of big data.

Source: OECD, 2016b.

The vast volumes of raw data, allow for analysis across multiple periods, taxpayers and tax domains. This enables administrations to plan their compliance, control and risk management. Relying more heavily on data analytics supports whole-of-government outcomes by sharing of insights and information between different sectors within government and between tax and other authorities all around the world. Such data are used for predictive analytics and simulations.

> Proactive analysis of taxpayer behaviour can help revenue bodies save time, money and effort during risk profiling. Risk profiling and scenario-based calculations are impacting the ability of administrations to target compliance interventions, whether at an industry segment or individual level, thereby contributing significantly to improving outcomes, reducing cost, and increasing levels of taxpayer satisfaction. (OECD, 2016b)

9.3. Access-Use Debate

In the legal domain, there is discussion about how to regulate Big Data. The current European legal regime is mostly focussed on the first phase, when the data are gathered. The European Union's (EU) General Data Protection Regulation (2016) is a case in point. Access-based regulation typically clashes fundamentally with the cure assumptions of Big Data.[3] Five examples are provided as follows:

1. *Purpose and purpose limitation*: First, the General Data Protection Regulation requires that data are gathered for a specific and concrete goal and that those data are not subsequently used for other, unrelated purposes. It specifies that personal data should be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes [] ('purpose limitation')".[4] It is obvious that this goes against the core assumption of Big Data, namely that data can be gathered without a specific goal, *inter alia* because data gathering and storage is so cheap. Often, the value and potential use of data are only determined after the data have been gathered. Another core assumption of Big Data is that data can always be reused, can always have a second life by reusing them for new purposes, by combining them with other data sets or by enriching them by scraping data from the internet. Reusing data for new purposes is a core aspect of Big Data, but fundamentally conflicts

[3] The GDPR only applies to personal data, Article 4.1 GDPR. There is discussion about the precise demarcation of this concept, but in general, most data are or can become personal data. *See also*: Article 29 Working Party, 'Opinion № 4/2007 on the concept of personal data', WP 136, 20 June 2007.

[4] Article 5.1.b GDPR. *See also*: Article 29 Working Party, Opinion 03/2013 on purpose limitation', WP 203, 02 April 2013.

with the purpose limitation principle in the General Data Protection Regulation (van der Sloot & van Schendel, 2016).

2. *Data minimization and storage limitation*: Second, the General Data Protection Regulation specifies that only those data can be gathered that are absolutely necessary for reaching the specific and predetermined purpose and that data should be deleted once that specific and predetermined purpose has been reached. It specifies that personal data should be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation')"[5] and should be "kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed ('storage limitation')".[6] Again, these principles go against the grain of new technological applications. The idea behind Big Data is rather to gather as many data as possible, to store them for as long as possible and to enrich data sets with as many different data sources.[7]
3. *Integrity and confidentiality*: Third, the General Data Protection Regulation specifies that the organization gathering data is responsible for upholding the legal principles. It should thus restrict the access to data to those people within the organization that need access for fulfilling their tasks, and should outright prevent third parties, such as hackers, from accessing those data. The Regulation specifies that data should be

   > processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures [] ('integrity and confidentiality').[8]

   Again, this idea goes against one of the essential characteristics of Big Data, namely that data should not be privatized and protected with barriers, but kept in an open and accessible way, sharing data, so that others can use those data for other purposes, develop new applications and pursue technological innovation.
4. *Transparency obligation and individual rights*: Fourth, the General Data Protection Regulation specifies that organizations gathering data about people must inform them, before gathering those data, about their identity, the fact that data are gathered, why,

[5] Article 5.1.c GDPR.

[6] Article 5.1.e GDPR.

[7] Necessity, proportionality and subsidiarity are legal concepts that are difficult to reconcile with Big Data technologies.

[8] Article 5.1.f GDPR.

how and to what end; if data are not directly obtained from citizens themselves, the organization must provide them with such information shortly after the moment of obtaining the data.[9] Vice versa, citizens can submit information requests to those organizations, to obtain information about whether data are stored about them, to what ends they are used and how they are processed.[10] In Big Data processes, however, it is often unclear to organizations about whom they have data, about which citizens data analysis makes predictions or from which social media profiles they have scraped personal data. Information flows often do not revolve around specific and identified individuals, but rather around larger groups of people; smart camera's film everyone walking in a certain street, intelligence agencies monitor the internet traffic of almost everyone, heat sensors register all bodies producing warmth, etc. In addition, there are simply so many organizations with so many databases in which data about one specific person can be contained that it becomes virtually impossible for a specific individual to know who those organizations are, to analyse whether or not they abide by all the legal principles and if they do not, to file an official complaint.

5. *Data flows*: A fifth and final example is that the General Data Protection Regulation specifies that personal data can be shared with organizations outside the EU only if they adhere *grosso modo* to the principles entailed in the General Data Protection Regulation. It holds:

> Any transfer of personal data which are undergoing processing or are intended for processing after transfer to a third country or to an international organisation shall take place only if, subject to the other provisions of this Regulation, the conditions laid down in this Chapter are complied with by the controller and processor, including for onward transfers of personal data from the third country or an international organisation to another third country or to another international organisation. All provisions in this Chapter shall be applied in order to ensure that the level of protection of natural persons guaranteed by this Regulation is not undermined.[11]

Again, Big Data technologies and the data flows that are part of them are almost by definition not bound by borders of countries or continents; rather, through the use of

[9] Articles 12, 13 and 14 GDPR.
[10] Article 15 GDPR.
[11] Article 44 GDPR.

> cloud computing and other means, data can be stored anywhere in the world and gathered from different citizens in different countries, for example by placing cookies and by scraping the internet.

In conclusion, the current legal paradigm primarily sets limits to the first phase of Big Data processes, in which data are gathered. It is clear that, although there are exceptions and exceptions to the exceptions to all the legal principles discussed above, in essence these legal principles clash with the basic philosophy behind Big Data. That is why a number of scholars and experts have called for a larger focus on the use of data instead, the third phase of Big Data processes. Three main arguments underlie such proposals.

1. First, they argue that the legal realm is simply outdated or out of touch with reality; such experts suggest that the General Data Protection Regulation, which has just recently been adopted and will only come into effect as of May 2018, is based on a 1990s understanding of data and technology. Whereas by the end of the last century it was still possible to set limits to the gathering of data, in the Big Data era access to information is simply a given. It is no longer possible to set restrictions on the gathering of data. In fact, these experts point to the actual state of affairs and suggest that the EU regulator is simply out of touch with reality and unaware of the phenomenon of Big Data.

   > It seems beyond doubt that this inadequate sense of realism also manifests itself in respect of privacy regulations. This inadequate sense of realism with regard to privacy regulations limits the chance of such regulations being embraced in anything more than a rhetorical manner, and indeed this has become the hallmark of existing privacy legislation. Frequently, there is a lack of any concrete development or real accountability with regard to the more universal claims and ambitions of these regulations. (Moerel & Prins, 2016)

2. Second, the argument is put forward that by setting limits to the first phase of Big Data, during which data are gathered, new innovations and the commercial exploitation of data by companies and governmental organizations are curtailed. They point to the value and potential of Big Data and suggest that even if it would be possible to enforce the current access-based legal principles and put an end to the mass collection of data by governmental organizations, companies and citizens alike, this would simply be undesirable because it would muffle the economic growth, technological progress and societal developments that are facilitated by Big Data (*see* for a discussion of use-based rules and regulation *inter alia*: Mundie, 2014; Horvitz & Mulligan, 2015). That is why proponents of a use-based regulation of Big Data suggest that the gathering of

personal data should no longer be restricted or that there should be substantially less hurdles in place for gathering personal data (*see* further: Culnan & Armstrong, 1999; Karjoth & Schunter, 2002).

3. Third, instead, the legal realm should be focussed on regulating the use of data, the third phase of Big Data processes, in which data and the insights gained from data analytics, are applied in practice and have a specific effect on citizens and society. It is the negative effects of data usage that should be curtailed, not the data gathering as such, such experts argue. When it is ensured that citizens and society alike benefit from Big Data processes, there is no reason to retain the access-based rules (*see* further: Moore, 2008). Consequently, data-driven innovation can flourish while privacy protection can actually be ameliorated, proponents of a use-based regulation of Big Data argue.

The General Data Protection Regulation currently only contains one provision that might be said to regulate the use of data. It holds:

> 1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her. 2. Paragraph 1 shall not apply if the decision: (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller; (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or (c) is based on the data subject's explicit consent. 3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision. 4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data [revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation], unless [the data subject has given explicit consent] or [processing is necessary for reasons of substantial public interest] and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.[12]

[12] Article 22 GDPR.

Although there are some exemptions and exemptions to the exemptions, again, the core of this legal principle is to prohibit automatic (non-human) decision-making based on profiling, when a citizen is significantly affected. The core idea behind this provision is that a general, abstract profile should not one-on-one be applied on a specific individual. Larger societal effects are left outside the realm of this provision, which experts suggest could be included.

In addition, there are other regimes that set limits to the use of data, such as when there are stigmatizing or discriminating effects. Reference can be made, *inter alia*, to Article 14 of the European Convention on Human Rights holding:

> The enjoyment of the rights and freedoms set forth in this Convention shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.[13]

Consequently, when profiling and data analytics are applied in practice by governmental agencies, they have the effect that people are discriminated against on the basis of race, sexual orientation, religious beliefs or other relevant aspects of their life, which might come into conflict with the non-discrimination principle (Moreau, 2013). In addition, there are national laws that prohibit certain uses of data; for example, the use of sensitive personal data for credit reporting or for making decisions about granting bank loans or setting insurance fees for citizens (Hoofnagle, 2013).

Still, in general, there are limited rules setting limitations on the use of data and the restrictions that are in place mostly focus on discrimination and sensitive personal data. Proponent of a use-based model of Big Data regulation suggest to elaborate and expand such rules and principles, setting boundary markers on how data can be used, by whom, when, for what reasons and in particular on what effects the use of data may and may not have. In this sense, a difference can be made between those that believe that the consequences of data usage should only be determined on an individual level – for example, when a specific person is denied a bank loan, is arrested based on discriminating assumptions or denied access to a university – and those that feel that the societal consequences should also be taken into account, referring to more structural and societal problems, such as the inequality and injustice that may follow from Big Data usage by companies and governmental organizations.[14]

[13] Convention for the Protection of Human Rights and Fundamental Freedoms, Rome, 4.XI.1950.

[14] To what extent privacy can be seen as a common or public good is a discussion point on its own. See among others: Regan (1995) and Allen (2011).

Proponents of an access-based model of regulation have three main arguments against adopting a use-based approach:

1. First, they believe that the proponents of a use-based approach are naïve; when there are no limits to gathering and storing data, they suggest, it will become unclear who has data, to what ends and how the data are used in practice. It is often unclear whether and to what extent decisions and applications in practice are in fact based on personal data, let alone who those data belong to. Analysing what effects applications have on people can be hard and very time-consuming and to what extent harms derive from data usage requires a tedious and meticulous process. If such matters of interpretation are not taken up by a governmental watchdog on structural level, this would mean that individuals would themselves need to analyse in how far certain data processes have a negative effect on their position and are themselves responsible for defending their interests through legal means. In reality, individuals are often powerless against the large and resourceful Big Data organizations.[15]
2. In addition, they suggest that the EU regulator is of course not out of touch with reality and is not unaware of Big Data technologies. Rather, the EU has signalled these developments and aims at stopping or curtailing those, among others, by setting very strict rules on when data can be gathered and how and by making clear in the General Data Protection Regulation that a violation of any of the five access-based rules discussed above can lead to a fine of up to €20 million or in the case of an commercial organization, up to 4% of the total worldwide annual turnover of the preceding financial year, whichever is higher.[16] This means that it is not entirely unrealistic that the amount of data currently gathered may be limited when the General Data Protection Regulation comes into effect.
3. Third, they suggest that a use-based approach is undesirable, because gathering data about persons as such is problematic and a violation of human freedom, not only when data are used and have negative consequences for specific people or society at large. Consequently, rules on the gathering of and the access to data should remain in place and be reinforced in the Big Data era, although perhaps, new use-based regulation can be added on top of the access-based principles. For example, Joris van Hoboken suggests:

> In a way, it almost appears as if the advocates of use-based regulation (as an alternative to the regulation of collection of personal data) believe that the concerns of individuals with respect to mere collection will miraculously disappear and that

[15] About to what extent it makes sense to grant individuals claim-rights, *see*: Van der Sloot (2017).
[16] Article 83 GDPR.

> European constitutional judges will eventually overturn their reasoning about data collection and fundamental rights. Considering the increased importance of personal data processing in all facets of society and the increased benefits as well as risks for data subjects, this seems both an unreasonable expectation as well as an undesirable way forward. Instead, it seems more reasonable to expect that more guarantees will be needed instead of less, to ensure the respect for information privacy and the continued trust in information technologies in a world in which any piece of data that is collected can end up being used for any purpose. (van Hoboken, 2016)

9.4. The Analysis of Data

Section 9.2 of this chapter argued that there is no clear and set definition of Big Data. Nevertheless, it is possible to discern three phases common to Big Data processes: gathering, analysing and using data. Section 9.3 argued that the current European legal paradigm, exemplified by the General Data Protection Regulation, is mostly focussed on setting rules and limits on the gathering and sharing of data, which is called an access-based approach to regulation. These rules fundamentally come into conflict with the core characteristics of Big Data. That is why it has been suggested to focus on the third phase of Big Data processes, which is called a use-based approach. These experts suggest to focus on the way in which data are used and the potential consequences data usage has for individuals or society at large.

Surprisingly, there is little attention to the second phase of Big Data processes, the phase in which data are analysed. In the legal realm, most instruments seem silent on this point. The exception to the rule may be the data accuracy principle in the General Data Protection Regulation, which suggests that personal data must be

> accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay (‘accuracy’).[17]

Consequently, Big Data organizations are required to ensure that the data they store are accurate and when data are in the possession of these organizations for longer periods of time, they should ensure that the data are kept up to date. This guarantees that the data analysed reflects reality and that the analysis and predictions are correct and reliable.

[17] Article 5.1.d GDPR.

In addition, there is little attention among academics and experts for the second phase of Big Data processes (for some first attempts on this point, *see*: Van der Sloot, 2013), although there is a growing interest for algorithmic decision-making and transparency (Diakopoulos, 2016; Kissell & Malamut, 2005; Maji, Roy & Biswas, 2002; Newell & Marabelli, 2015). Still, it is this phase where perhaps most of the mistakes are made, where the negative consequences for individuals and society derive from and where introducing standards and rules could be most beneficial. It is impossible to give a full overview of all the errors that are commonly made in Big Data processes, but ten common obstacles are discussed as follows.[18]

1. *Data are not neutral*: Big Data processes involve statistical analysis. As everyone knows that has studied statistics, designing a proper research methodology, collecting reliable data and finding statistically relevant and significant correlations is hard. A first requirement to ensure reliable statistical outcomes is securing the reliability of the data and the data set and to correct potential biases. First, the data must be representative of the actual situation. Suppose of the customers of a certain company, 34% are male, but the data set contains data about 9,000 women and only about 1,000 men, this should be corrected in the data model. In the enthusiasm for Big Data, this simple step is often forgotten, not in the last place because organizations are not always aware of how the data were gathered, what biases existed in the research methodology and what would be a proper representation of reality. This problem is aggravated when scraping data from the internet, such as from Facebook, Twitter or other social media. Not only is it impossible to know what biases exist in the data and how to correct those, in addition, the way in which the platforms are designed also influence which data are posted by users and how. Designing a proper research methodology for gathering reliable data requires time and effort, because the way in which questions are posed to people can have a large influence on their answers.[19] To provide a final example of the problem of biased data sets in Big Data processes is what is called the feedback loop. Take the police. The police traditionally surveys more in areas where many conflicts and problems arise. In the Netherlands, this would be the Bijlmer in Amsterdam, the Schilderswijk in The Hague and Poelenburg in Zaandam. A substantial part of the population in these neighbourhoods are immigrants or people that have an immigrant background. Consequently, these groups have an above-average representation in the police databases. Subsequently, when the police decides where it should survey and deploy units, data analytics will suggest to focus on these

[18] These points are translated from Van der Sloot (forthcoming).

[19] Add to this the point of emotion experimentation: https://www.theatlantic.com/technology/archive/2014/06/everything-we-know-about-facebooks-secret-mood-manipulation-experiment/373648/

neighbourhoods and people with these backgrounds, which will lead to an even higher number of data points about these areas and the groups living in those areas, etc.

2. *Updating data is not neutral*: A second problem is that Big Data analytics is often based on incorrect data. Although the catch phrase of Big Data gurus is that Big Data can work with messy and incorrect data, 'quantity over quality', reality often proves different. Inadequate data lead to inadequate analytical results. In addition, data are often analysed when they are outdated. Out of the feeling that data can always have a second life and be reused for new purposes, many organizations use the databases already in their possession for Big Data applications. Making predictions on the basis of old data obviously predicts the past. When organizations update the databases, a number of things go wrong. Most importantly, the metadata about the data set – how the data have been gathered, when, why and by whom – is often non-existent; this makes it impossible to complement the outdated data set with new data using the same methodology. Consequently, different data, with different biases are incorporated in the same database, which makes it almost impossible to correct those biases, which has the effect that the comparison between the data is inaccurate. Comparing data of different periods with the same bias may signal significant differences over time; when the data sets compared have different research designs, it becomes impossible to exclude the possibility that the differences found in the data sets are a result of the difference in methodology. More in general, a number of different factors may have an effect on why data about different time frames may be different, such as difference in legal and societal norms, difference in population, difference in the world economy, difference in environments, etc. These circumstances are often simply excluded from the research design.

3. *Categorizing data is not neutral*: Although again, a catch phrase of Big Data analytics is that it can work with messy and unstructured data, reality so far proves to be less bright. Categories are essential to understanding data, while at the same time, categories are non-neutral. Should a police database also include data about race, gender or religious background? Take only the most basic of categories 'Are you either male [] or female []?', which has been questioned recently, *inter alia,* by people that feel they are neither and by people that affiliate with both genders. Choosing categories is a non-neutral endeavour. In addition, setting boundaries on who falls within which category often proves to be quintessential for the outcome of the analysis. This is a problem in Big Data processes in itself, because the categorization in databases is often a result of usability (the databases in the possession of organizations were traditionally used by employees, board members and others) and not on reliability. This problem is aggravated when different databases are merged; even when the databases contain data on the same subject matter, it can be difficult to make choices on the

categorization of the merged database. A well-known example is that in one data set, 'young adults' are persons between the age of 16 to 28, whereas in another data set, the same term is used for persons between 18 to 26. Which of these two categories is used can have a large impact on the results. Every data set has its own background and there are often reasons for specific demarcations. The first database concerns, for example, the moment when young people first have sex (with the age of 16 as the legal age for having sex with a person), while the second data set may be about the intake of alcohol by young people (with the age of 18 years as the threshold for legally buying alcohol). These categories can therefore not be changed without significantly perverting the reliability of the data set, whereas such practices are not uncommon to Big Data processes.

4. *Algorithms are not neutral*: In addition, algorithms, the computer models used to analyse the data, themselves are not neutral. They are based on decision trees, which attach weight to certain factors more than to others. A good algorithm makes more accurate predictions than a bad algorithm, but both are based on assumptions that are never always true. It is a decision tree that includes certain factors, but ignores others, attaches weight to factors and draws conclusions from the different weights. Although programmers are themselves often aware of the assumptions and biases in the algorithms and the need to correct the findings when using or applying the results gained through data analysis, this awareness is often lost on a managerial or board room level. In addition, there is the problem of blind spots. A well-known example is that predictive algorithms initially simply ignored the likelihood that women would commit terrorist attacks, making black widows the ideal perpetrators. The other way around, algorithms can have a bias towards certain ethnic groups, so that too much emphasis is put on those, making the data-based application ineffective and inaccurate.[20]

5. *Queries are not neutral*: As has been discussed, the data and the categories are not neutral, but also the queries run on those data should not be conceived as neutral. Suppose the police runs a database query searching for Muslims that are likely to commit a burglary in the coming month. Suppose the results show that there are five people that have a high probability of committing such crime. When the police starts monitoring those people, it can point to credible intelligence that these people may commit a crime. Although it is not unreasonable that the police should monitor those people’s behaviour, the underlying research query is clearly biased.[21] This example is extreme, namely based on intentional discrimination. More in

[20] Algorithms are often also trained on data already in the possession of organizations. This means that the algorithm learns to recognize faces, for example, primarily on photos from white people, which may have a consequence for recognizing the faces of people with darker skin. Such biases are also well-known in medical research, which are tested mostly on young white males.

[21] A fact that is difficult to address through the use-based regulation of Big Data.

general, however, every research query is biased as it is based on assumptions, a certain phrasing and potential expectations on the outcome. In addition, discrimination may not only follow from direct and intentional actions, but also as an indirect result (*see* for a discussion on redlining: Squires, 2003). A query on postal code may have an important bias, as certain areas have a larger immigrant population than others, etc. In Big Data research, such queries are often put in the hand of managers, who more often than not do not have a clear understanding of the implicit assumptions in their research questions or the biases in the queries.

6. *Predictions are not facts*: Big Data analytics typically revolve around predictions and probabilities. It can be found, through data analytics, that there is a 70% likelihood that men in possession of a red car and with right-wing political beliefs will read the *Wall Street Journal*, or that there is a 23% chance that certain types of bridges will collapse when there is an earthquake, or a 44% chance that extreme obesity will be the first cause of death in Europe and the United States of America. There are two common mistakes in this respect. First, when using and applying the insights gained from data analytics, the predictions tend to be taken absolute, while the outcomes are probabilistic. Second, obviously, data analytics does not predict anything about a specific individual, car or house. It predicts something about the group. This is one of the reasons why doctors are often hesitant to predict the progression of certain diseases. The fact that 70% of the people recover from the disease says little about an individual patient. With Big Data processes, however, it is not uncommon to apply general correlations on specific individuals, such as with terrorism or crime prevention.
7. *False positives and negatives*: As predictions are not facts, there are always problems in terms of false positives and false negatives. When there is a false positive, a prediction is made about something or someone that is ultimately not correct. This can sometimes be harmless, for example when someone is shown an advertisement of a product in which he is not actually interested. It can already be a waste of time when a bridge is predicted to be in need of repair, while it follows from the subsequent inspection that this is not the case. False positives are particularly problematic in the medical sector, when a patient is predicted to attract a particular disease, while that disease does not manifest itself at all. It is also problematic if a person is suspected and accused of a crime, while he turns out to be innocent. It can have a big impact on people if they are unjustly suspected of a crime or wrongly predicted to be ill in the future, and family life can be disrupted as a consequence. In the case of false negatives, there is the opposite problem. Again, little harm is done when someone does not see an advertisement of a product in which he is interested or if the automatic refrigerator has not ordered a new bottle of milk. More problematic is it when a person is not spotted as a

potential terrorist, while he or she prepares for an attack, or when a disease is not found in time, when the medical institution relies on Big Data analytics. One of the problems is that there are no thresholds in place for how high the number of false negatives and positives can be, so that organizations tend to spend little effort in marginalizing these false outcomes; at the same time, they often take the consequences of these false predictions as a given, as an integral part of doing Big Data analysis.

8. *Correlation is not causality*: What is a reoccurring problem with Big Data analytics is a confusion of statistical correlations and causality. The fact that someone places felt pads under the legs of his chairs and tables can have a predictive value as to whether this person will repay his loan. However, it is not because he places felt pads under the legs of his chairs and tables that he repays his loan. The same goes for religious, cultural and ethnic background in relation to crime; although there may be a statistical correlation, this says nothing about causality. There is a famous anecdote about how this almost went wrong in the United States of America. The story goes that the governor of an American state saw with dismay that the children attending school often performed poorly and did not attend university. A large data-driven study was carried out into the school performance of children and it appeared that one of the factors with the greatest predictive value for school performance was the number of books in the house where the children grew up. The governor then decided to draw up a book plan; all the households in which children grew up had to be sent books to promote the school performance of children. Only at the last moment would this plan have been cut off, because the causal relationship was non-existent. It is not that children get smart because there are many books in their home, it is much more likely that, for example, highly educated parents both possess many books and stimulate and support their children in their school achievements.

9. *Experimentation*: As a penultimate example of what might go wrong when analysing the data, it often occurs that large and general predictions are based on too small data sets (under the pretext of Big Data). This especially occurs in smaller organizations, such as schools, archives and retailers, who are keen on applying 'Big Data analytics' in their organization, but only have a limited data set to work with. It is not uncommon that large policy decisions and strategic plans are based on a data set with an *n* lower than 100. For example, some secondary schools base their admission policy on the data they have collected from one or two subsequent years. They have analysed from which district of the city the pupils came, which school advice they had and for example whether they are male or female and connect this to how they performed at high school. On the basis of data analytics applied to data from one or two years, admission policies are designed. Obviously, such small samples cannot result in reliable

statistical predictions, as every year is unique. Only with larger data sets can general patterns be discerned in a reliable manner.

10. *Falsification:* As a final example of what often goes wrong with data analytics is that there is hardly any falsification of the results. A good example is predictive policing, about which there is little scientific evidence supporting its supposed efficacy. Rather, a number of police forces have stopped working with predictive policing out of a lack of results. Still, some police forces believe that in their case, predictive policing is effective, pointing to a decline of crime rates after the introduction of predictive policing. The causal relationship remains unproven. Comparative research often shows, for example, that in the same period in other cities, crime rates have also dropped, without deploying predictive policing. This may be due to, for example, general developments on which Big Data has no influence. Alternatively, positive outcomes may be due to the fact that predictive policing is applied as part of a broader strategy, for which all kinds of means are used. To assess the reliability of the results, as a minimum, the following steps should be taken: a baseline measurement, setting the goal of applying Big Data analytics, monitoring the results, analysing whether there is a positive impact vis-à-vis the baseline measurement, analysing in how far this can be attributed to the deployment of Big Data analytics, and finally a falsification of the potential positive results.

9.5. Building Blocks

Section 9.2 of this chapter argued that Big Data processes can be divided in three phases: gathering data, analysing data and using data. Section 9.3 argued that the current European legal paradigm mostly focussed on setting rules and limits on the gathering of data, while it has been suggested to lay emphasis on the regulation of the third phase of Big Data processes, in which data are used. Section 9.4 suggested that it is particularly in the second phase, in which data are analysed, that there are glaring errors and mistakes. Examples given were the biases in data sets, in algorithms and in search queries, the confusion of correlation for causality and the lack of validation and falsification. This has the effect that Big Data analytics become less reliable, making governmental organizations that rely on the results less trustworthy, which can have a negative impact on the attitude of citizens towards the law and legal obligations.

The reason why the defenders of an access-based regulation of data and the defenders of an use-based regulation of data have both mostly ignored this phase is because privacy and data protection are commonly framed as rights protecting individual interests. The scope of data protection instruments is determined by the term ‘personal data’, which is defined as any data that can be used to identify a person. The defenders of an access-based regulation focus on the moment at which data are gathered from a person. Thus, the link to the personal interest is still evident. The proponents of an use-based regulation focus on the moment at which the insights gained from the data analytics are applied in practice, and thus as a direct and concrete effect on individuals and groups. The phase in between, when the data are analysed and patterns and profiles are distilled, is perhaps the core phase of Big Data processes, but no direct individual interest is linked to this phase. There is no harm as such in analysing incorrect data, there is no harm in using biased algorithms as such or in applying methodologically unsound data analytics (Figure 9.2).

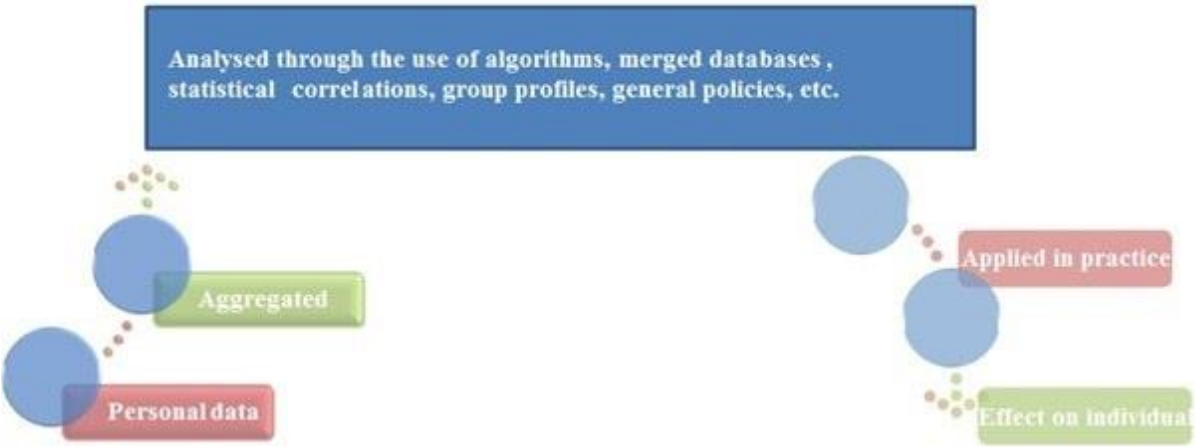

Analysed through the use of algorithms, merged databases , statistical correlations, group profiles, general policies, etc.
Aggregated
Personal data
Applied in practice
Effect on individual

Figure 9.2.

The point is that the data that are processed in Big Data initiatives often do not directly identify a person, but are gathered, assessed and used on a general, aggregated or group level. For example, they may be used to adopt policies on the basis of zip codes, income levels or any other general criterion. Thus, these data do not directly identify a person, and consequently fall outside the scope of the data protection regulations, even though they may affect the data subject as being part of a specific group. Of course one could focus on the initial moment, when personal data are gathered and not yet aggregated, but this may only concern the split second which it takes to aggregate data. The same goes for the moment at which group profiles are applied and used to affect a specific person (if this can be determined). The use of data only concerns the very end of the data process. By focussing on the individual, her interests and her rights, one loses from sight the larger part of the data processing scheme and the general issues concerned with it. Focussing on the second phase of Big Data – when the data are analysed and patterns and profiles are gained – has the benefit of addressing one of the core challenges of now mostly left unregulated.

Consequently, the current framework could be ameliorated by introducing rules and standards for analysing data. This can be done through legal means, but other forms of regulation, such as informal control, social norms, self-regulation and third party regulation may equally serve the purpose.

Although there are a number of laws that contain standards and principles that could be a source of inspiration, most focus on the fairness of the outcome, rather than the process itself. There is one exception in the General Data Protection Regulation, already mentioned, which requires the quality of the data. In addition, in the recitals to the General Data Protection Regulation, there is mentioning of statistical principles and procedures that should be respected.

> In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised [].[22]

Consequently, it might be worthwhile looking to statistical principles and procedures for ameliorating Big Data regulation and laying down rules for the second phase of Big Data processes. In the Treaty on the Functioning of the EU, such principles are already embedded. Article 338 specifies:

> 1. Without prejudice to Article 5 of the Protocol on the Statute of the European System of Central Banks and of the European Central Bank, the European Parliament and the Council, acting in accordance with the ordinary legislative procedure, shall adopt measures for the production of statistics where necessary for the performance of the activities of the Union. 2. The production of Union statistics shall conform to impartiality, reliability, objectivity, scientific independence, cost-effectiveness and statistical confidentiality; it shall not entail excessive burdens on economic operators.[23]

Among others, these principles are elaborated on in the Regulation on European Statistics,**[24]** of which Article 2 holds that the following principles should be respected:

[22] Recital 71 GDPR.

[23] *See* http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012E/TXT&from=EN

[24] Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics and repealing Regulation (EC, Euratom) No 1101/2008 of the European Parliament and of the Council on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities, Council Regulation (EC) No 322/97 on Community Statistics, and Council Decision 89/382/EEC, Euratom establishing a Committee on the Statistical Programmes of the European Communities (Text with relevance for the EEA and for Switzerland).

- *Professional independence*: statistics must be developed, produced and disseminated in an independent manner, particularly as regards the selection of techniques, definitions, methodologies and sources to be used, and the timing and content of all forms of dissemination, free from any pressures from political or interest groups or from Community or national authorities, without prejudice to institutional settings, such as Community or national institutional or budgetary provisions or definitions of statistical needs;
- *Impartiality*: must be developed, produced and disseminated in a neutral manner, and that all users must be given equal treatment;
- *Objectivity*: statistics must be developed, produced and disseminated in a systematic, reliable and unbiased manner; it implies the use of professional and ethical standards, and that the policies and practices followed are transparent to users and survey respondents;
- *Reliability*: statistics must measure as faithfully, accurately and consistently as possible the reality that they are designed to represent and implying that scientific criteria are used for the selection of sources, methods and procedures;
- *Statistical confidentiality*: the protection of confidential data related to single statistical units, which are obtained directly for statistical purposes or indirectly from administrative or other sources and implying the prohibition of use for non-statistical purposes of the data obtained and of their unlawful disclosure;
- *Cost effectiveness*: the costs of producing statistics must be in proportion to the importance of the results and the benefits sought, that resources must be optimally used and the response burden minimized. The information requested shall, where possible, be readily extractable from available records or sources.

In addition, Article 12 of the Regulation elaborates that in order to ensure statistical quality, the following criteria should be adhered to:

- *Relevance*: the degree to which statistics meet current and potential needs of the users;
- *Accuracy*: the closeness of estimates to the unknown true values;
- *Timeliness*: the period between the availability of the information and the event or phenomenon it describes;
- *Punctuality*: the delay between the date of the release of the data and the target date (the date by which the data should have been delivered);
- *Accessibility and Clarity*: the conditions and modalities by which users can obtain, use and interpret data;

- *Comparability*: the measurement of the impact of differences in applied statistical concepts, measurement tools and procedures where statistics are compared between geographical areas, sectoral domains or over time;
- *Coherence*: the adequacy of the data to be reliably combined in different ways and for various uses.

Additionally, reference can be made to the United Nation's General Assembly, which has adopted ten fundamental principles of statistics.[25] Referring to the critical role of high-quality official statistical information in analysis and informed policy decision-making and stressing that

> the essential trust of the public in the integrity of official statistical systems and confidence in statistics depend to a large extent on respect for the fundamental values and principles that are the basis of any society seeking to understand itself and respect the rights of its members, and in this context that professional independence and accountability of statistical agencies are crucial,

it specifies the following ten principles:[26]

1. *Utility*: Official statistics provide an indispensable element in the information system of a democratic society, serving the government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled with and made available on an impartial basis by official statistical agencies to honour citizens' entitlement to public information.
2. *Trust*: To retain trust in official statistics, the statistical agencies need to decide according to strictly professional considerations, including scientific principles and professional ethics, on the methods and procedures for the collection, processing, storage and presentation of statistical data.
3. *Scientific reliability*: To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics.
4. *Educational role*: The statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics.

[25] *See also*: https://unstats.un.org/unsd/dnss/gp/FP-Rev2013-E.pdf

[26] Resolution adopted by the General Assembly on 29 January 2014 [without reference to a Main Committee (A/68/L.36 and Add.1)] 68/261. Fundamental Principles of Official Statistics.

5. *Quality, timeliness, costs and burden-sharing*: Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies are to choose the source with regard to quality, timeliness, costs and the burden on respondents.
6. *Confidentiality*: Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.
7. *Transparency*: The laws, regulations and measures under which the statistical systems operate are to be made public.
8. *Coordination*: Coordination among statistical agencies within countries is essential to achieve consistency and efficiency in the statistical system.
9. *Consistency*: The use by statistical agencies in each country of international concepts, classifications and methods promotes the consistency and efficiency of statistical systems at all official levels.
10. *Cooperation*: Bilateral and multilateral cooperation in statistics contributes to the improvement of systems of official statistics in all countries.

As a final source of inspiration, reference can be made to the European Statics Code of Practice, adopted in 2011 by the European Statistical System Committee, which incorporate, as far as relevant here, the following principles (paraphrased):[27]

1. *Professional environment*:
   a. *Independence:* Independence from political and other external interference in developing, producing and disseminating statistics is specified in law and assured for other
   b. *Access:* Access to policy authorities and administrative public bodies.
   c. *Competence*: The capacity of the head and employees of the organizations is beyond dispute and other arguments than their competence should play no role in their appointment.
   d. *Transparency:* There is openness on the statistical work programs.
   e. *Clarity:* Statistical releases are clearly distinguished and issued separately from political/policy statements.
2. *Mandate and recourses*:
   a. *Mandate:* Authorities should have a mandate to gathered and process data for statistical analysis.

[27] The European Statistics Code of Practice, http://ec.europa.eu/eurostat/documents/64157/4392716/Revised_CoP_Nov_2017.pdf

    b. *Recourses:* Staff, financial, and computing resources, adequate both in magnitude and in quality, are available to meet current statistical needs.

3. *Quality oversight*: Authorities should monitor the quality of their statistical analysis, evaluate their work and the procedures in place to guarantee quality and external experts are consulted when appropriate.
4. *Privacy*: the data are kept safely and confidentially, both organizationally and technically.
5. *Objectivity*:
    a. *Objective compilation:* Statistics are compiled on an objective basis determined by statistical considerations.
    b. *Choices based on statistical considerations:* Choices of sources and statistical methods as well as decisions about the dissemination of statistics are informed by statistical considerations.
    c. *Correction of errors:* Errors discovered in published statistics are corrected at the earliest possible date and publicized.
    d. *Transparency:* Information on the methods and procedures used is publicly available.
    e. *Notice of changes:* Advance notice is given on major revisions or changes in methodologies
    f. *Objective and non-partisan:* Statistical releases and statements made in press conferences are objective and non-partisan.
6. *Quality*:
    a. *Consistency:* Procedures are in place to ensure that standard concepts, definitions and classifications are consistently applied throughout the statistical authority.
    b. *Evaluation:* The business register and the frame for population surveys are regularly evaluated and adjusted if necessary in order to ensure high quality.
    c. *Concordance:* Detailed concordance exists between national classifications systems and the corresponding European systems.
    d. *Relevant expertise:* Graduates in the relevant academic disciplines are recruited.
    e. *Relevant training:* Statistical authorities implement a policy of continuous vocational training for their staff.
    f. *Cooperation with scientific community:* Cooperation with the scientific community is organized to improve methodology, the effectiveness of the methods implemented and to promote better tools when feasible.
7. *Validation*:

    a. *Prior testing:* In the case of statistical surveys, questionnaires are systematically tested prior to the data collection.
    b. *Reviewing:* Survey designs, sample selections and estimation methods are well based and regularly reviewed and revised as required
    c. *Monitoring:* Data collection, data entry, and coding are routinely monitored and revised as required.
    d. *Editing:* Appropriate editing and imputation methods are used and regularly reviewed, revised or updated as required.
    e. *Designing:* Statistical authorities are involved in the design of administrative data in order to make administrative data more suitable for statistical purposes.

8. *Reporting burden*:
    a. *Proportionality:* Reporting burden is proportionate to the needs of the users and is not excessive for respondents.
    b. *Necessity:* The range and detail of the demands is limited to what is absolutely necessary.
    c. *Egality:* The reporting burden is spread as widely as possible over survey populations.
    d. *Availability:* The information sought from businesses is, as far as possible, readily available from their accounts and electronic means are used where possible to facilitate its return.
    e. *No duplicity:* Administrative sources are used whenever possible to avoid duplicating requests for information.
    f. *Generality:* Data sharing within statistical authorities is generalized in order to avoid multiplication of surveys.
    g. *Linkability:* Statistical authorities promote measures that enable the linking of data sources in order to reduce reporting burden.
9. *Reality and comparability*:
    a. *Validation:* Source data, intermediate results and statistical outputs are regularly assessed and validated.
    b. *Documentation:* Sampling errors and non-sampling errors are measured and systematically documented according to the European standards.
    c. *Analysed:* Revisions are regularly analysed in order to improve statistical processes.
    d. *Coherent:* Statistics are internally coherent and consistent (*i.e.* arithmetic and accounting identities observed).
    e. *Comparable:* Statistics are comparable over a reasonable period of time.

    f. *Standardification:* Statistics are compiled on the basis of common standards with respect to scope, definitions, units and classifications in the different surveys and sources.
    g. *Comparability:* Statistics from the different sources and of different periodicity are compared and reconciled.
    h. *Exchange:* Cross-national comparability of the data is ensured through periodical exchanges between the statistical systems.

10. *Accountability*:
    a. *Metadata:* Statistics and the corresponding metadata are presented, and archived, in a form that facilitates proper interpretation and meaningful comparisons.
    b. *Dissemination:* Dissemination services use modern information and communication technology and, if appropriate, traditional hard copy.
    c. *Transparency:* Custom-designed analyses are provided when feasible and the public is informed.
    d. *Microdata:* Access to microdata is allowed for research purposes and is subject to specific rules or protocols.
    e. *Metadata:* Metadata are documented according to standardized metadata systems.
    f. *Information:* Users are kept informed about the methodology of statistical processes including the use of administrative. Users are kept informed about the quality of statistical outputs with respect to the quality criteria.

9.6. Conclusion

The reliability and trustworthiness of governmental organizations is not only important to ensure legitimacy, but also their effectiveness. People that feel that the government is unreliable, unpredictable and not trustworthy will typically have a more relaxed approach towards the law and will be less willing to cooperate with official institutions. When tax authorities rely on incorrect data, on biased algorithms or faulty methodology, the man-hours and money are spent inefficiently. Hence, one of the promises of Big Data – to increase efficiency and effectiveness of policies – will not or only marginally be realized. In addition, when tax authorities make incorrect assumptions about citizens and companies and approach them in a biased way, trust will wither just like trust of citizens in the police has declined when using biased predictive policing tools.[28]

Introducing standards for analysing data would optimize the analysis of data and hence increase the reliability of data analytics. This would in its turn have a positive effect on the

[28] *See* Thomas (2016).

trustworthiness and reliability of government agencies relying on Big Data usage and the confidence of citizens in their government and the exercise of power by governmental agencies. The principles discussed in the previous section could be taken as a starting point for regulating the second phase of Big Data processes, in which data are categorized and analysed.

Regulation can be either through law, through co- or self-regulation, by promoting awareness or through other means. There may be two reasons why focussing solely on black letter law provisions may not be effective in the age of Big Data.

First, the current paradigm places its bets mainly on the legal regulation of rights and obligations – black letter law. Yet it is increasingly questionable whether and to what extent this form of regulation still suffices in the Big Data era. That has to do with a number of issues. First, data processing is increasingly transnational. This implies that more and more agreements need to be made between different states and organizations in different jurisdictions. Hard legal rules are often difficult to agree upon due to the difference in traditions and legal systems. Furthermore, rapidly changing technology has the effect that specific legal provisions can easily be circumvented and that unforeseen problems and challenges may arise. And, as discussed, many of the problems arising from Big Data practices are social and societal. It is questionable whether those concerns should be dealt with fully within the juridical discourse. It could be promising to regulate Big Data processes additionally through forms of soft law and ethical standards, such as duties of care and codes of conduct. The underlying normative principles and values to be guaranteed in Big Data processes remain relatively stable. One could also look to other sectors for inspiration, for example the idea of installing ethical oversight committees, such as is a common practice in the medical sector. An interdisciplinary group of experts, consisting for instance of lawyers, ethicists, engineers and practitioners, could assess specific plans, policies and experiments.

Second, the current regulatory regime is based on numerous categorizations, labels and distinctions. For example, distinctions can be made between the offline and online, between the analogue and digital environment, between the protection of privacy in the private and in the public domain, between different nations and jurisdictions, between times of war and times of peace, between the powers and capacities of organizations in the private sector and in the public sector, and between different organizations in the public sector (for example in relation to which data they may gather, how they might use them and for what purposes; the intelligence agencies have broader powers to process data than the police, and the police has broader powers than the social services). In the Big Data era, however, the world is becoming increasingly fluid. Although the rights of citizens are currently linked mainly to physical objects such as the body and the home, and certain forms of communication such as the secrecy of correspondence, the Big Data era requires that one's digital identity, internet communications and privacy in the public domain be protected equally. Likewise, in

Big Data processes, data streams increasingly circulate between the private and the public sector and between different governmental agencies. Future regulation will need to standardize the rules applicable to those different sectors.

Consequently, it could be worthwhile to adopt a code of conduct, in which tax authorities affirm that they will endeavour to adopt a number of principles which guide Big Data processes in the second phase, in which data are analysed and patterns and profiles are distilled. Such code of conduct can be subdivided into organizational principles, statistical principles, rules on transparency and oversight, and rules on comparability and compatibility. In addition, a number of principles can be developed for Big Data specific contexts.

Organizational principles: First, there should be rules on the independence of the staff and the organization – they cannot be influenced. Second, there should be rules on the neutrality of the staff and organization – their personal or political opinion cannot have an influence on the data analysis. Third, there should be rules on the quality of the staff and the organization – people should have a proper training and acquire relevant insights and knowledge through continuous training. Fourth, there should be enough staff and tools available – high-quality research depends on the time and means to conduct research. Fifth, advice from and cooperation with experts is sought to ensure the reliability of the process.

Statistical principles: First, the gathering of data must be executed in a neutral and objective manner. Second, updating data must be done in a neutral and objective manner and accord to the original research design. Third, categorization of data must be done in a neutral and objective manner. Fourth, research queries must be made in a neutral and objective manner. Fifth, the algorithms used to analyse the data must be objective and neutral.

Transparency and oversight: First, the methods of research and analysis should be recorded. Second, those methods should be made public. Third, any changes in the methods should be recorded and made public; errors and biases should be corrected and made public. Fourth, internal audits should be conducted to analyse the correctness and efficacy of the methods, both prior, during and after the analysis of data. Fifth, external audits by experts or other organizations should be allowed and promoted – prior, during and after the analysis of data.

Comparability and compatibility: First, metadata on the database and analysis process should be kept. Second, gathering, classification and categorizing data should follow the rules and procedures commonly used by other organizations. Third, research methods and tools should align to those commonly used by other organizations. Fourth, there should be an equal spread in data about parts of the population. Fifth, when databases are integrated or merged, categorization and analysis should ensure the reliability of the merged data set and the data analysis following from it.

Big Data specific rules: First, set thresholds for the margin of error (false positives and false negatives) allowed for analysis; this may differ per organizations and context. Second, avoid feedback loops by verifying the design of the databases and the methods for gathering those data. Third, prevent decisions bases on one domain of life for aspects of fully different domains of life. Fourth, never treat correlation as causality, never treat predictions as facts. Fifth, always take the following steps: a baseline measurement, setting the goal of applying Big Data analytics, monitoring the results, analysing whether there is a positive impact vis-à-vis the baseline measurement, analysing in how far this can be attributed to the deployment of Big Data analytics, and finally a falsification of the potential positive results.

Whether Big Data will deliver on all its promises remains to be seen. What is certain is that if data analytics remains at the level of quality it is now, it will not. Standards need to be adopted that ameliorate the analysis of data. Currently, Big Data can be described as statistics for non-statisticians. This chapter has made a first effort to develop rules and principles that could guide the second phase of Big Data analytics, in which data are analysed. Although currently, there is much attention for the gathering and the use of data – that is, the moments at which there is still a direct link to the individual – it is the second phase that is perhaps most determinative for the Big Data process and in need of the greatest improvement. An international code of conduct might provide a first attempt at such improvement.

References

Algemene Rekenkamer. (2017). Tussenstand Investeringsagenda Belastingdienst. Available from https://www.rekenkamer.nl/binaries/rekenkamer/documenten/rapporten/2017/10/11/tussenstand-investeringsagenda-belastingdienst/Rapport+Tussenstand+Investeringsagenda+Belastingdienst+WR.pdf

Allen, A. L. (2011). *Unpopular Privacy: What Must We Hide?* New York: Oxford University Press.

Article 29 Working Party, 'Opinion Nº 4/2007 on the concept of personal data', WP 136, 20 june 2007.

Article 29 Working Party, Opinion 03/2013 on purpose limitation', WP 203, 02 April 2013.

Beck, U. (1992). *Risk Society Towards a New Modernity*. London: Sage Publications.

Bollier, D. (2010). The Promise and Peril of Big Data. www.emc.com/collateral/analyst-reports/10334-ar-promise-peril-of-big-data.pdf

Boyd, D., & Crawford, K. (2011). Six Provocations for Big Data. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*, 662-679.

Consolidated version of the treaty on the functioning of the European Union. Available from http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012E/TXT&from=EN

Convention for the Protection of Human Rights and Fundamental Freedoms, Rome, 4.XI.1950.

Craig, T., & Ludloff, M. E. (2011*). Privacy and Big Data: The Players, Regulators, and Stakeholders.* Sebastopol: O'Reilly Media.

Crain, M. (2018). The limits of transparency: Data brokers and commodification. *New Media & Society, 20*, 88-104.

Crawford, K., & Schultz, J. (2014). Big data and due process: Toward a framework to redress predictive privacy harms. *Boston College Law Review, 55*, 93-128.

Culnan, M. J., & Armstrong, P. K. (1999). Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. *Organization Science, 10*, 104-115.

Custers, B. H. M. (2004). *The Power of Knowledge; Ethical, Legal, and Technological Aspects of Data Mining and Group Profiling in Epidemiology*. Nijmegen: Wolf Legal Publishers.

Davis, K., & Patterson, D. (2012). *Ethics of Big Data: Balancing Risk and Innovation*. Sebastopol: O' Reilly Media. Available from www.commit-nl.nl/sites/default/files/Ethics%20of%20Big%20Data_0.pdf

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM, 59*(2), 56-62.

Dijcks, J. P. (2012). Oracle: Big Data for the Enterprise. Oracle White Paper. Available from www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf

Dusseault, P. L. (2013). Privacy and social media in the Age of Big Data: Report of the Standing Committee on Access to Information, Privacy and Ethics. http://www.parl.gc.ca/content/hoc/Committee/411/ETHI/Reports/RP6094136/ethirp05/ethirp05-e.pdf.

The European Statistics Code of Practice. (2017). For the National Statistical Authorities and the European Union Statistical Authority. Available from http://ec.europa.eu/eurostat/documents/64157/4392716/Revised_CoP_Nov_2017.pdf

Floridi, L. (2012). Big data and their epistemological challenge. *Philosophy and Technology*, *25*(4), 435-437.

Laney, D. (2012). 3D Data Management Controlling Data Volume Velocity and Variety. Available from https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Hildebrandt, M. (2015). *Smart Technologies and the End(s) of Law. Novel Entanglements of Law and Technology.* Cheltenham: Edward Elgar.

van Hoboken, J. (2016). From collection to use in privacy regulation? A forward-looking comparison of European and us frameworks for personal data processing. In B. van der Sloot, D. Broeders, & E. Schrijvers (Eds.). *Exploring the Boundaries of Big Data* (p. 249). Amsterdam: Amsterdam University Press.

Hoofnagle, C. J. (2013). How the Fair Credit Reporting Act Regulates Big Data. Available from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2432955

Hopkins, B., & Evelson, B. (2011). Expand Your Digital Horizon with Big Data. Forrester. Available from www.asterdata.com/newsletter-images/30-04-2012/resources/Forrester_Expand_Your_Digital_Horiz.pdf

Horvitz, E., & Mulligan, D. (2015). Data, privacy, and the greater good. *Science, 349*(6245).

IBM. Big Data Analytics. Available from www-01.ibm.com/software/data/bigdata/what-is-big-data.html

Karjoth, G., & Schunter, M. (2002). A Privacy Policy Model for Enterprises. Computer Security Foundations Workshop, IEEE.

Kerr, I., & Earle, J. (2013). Prediction, preemption, presumption. How big data threatens big picture privacy. *Stanford Law Review Online, 66*, 65-72.

Kissell, R., & Malamut, R. (2005). Algorithmic decision-making framework. *The Journal of Trading, 1*, 12-21.

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences.* Los Angeles: Sage.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: Traps in big data analysis. *Science, 343*, 1203-1205.

Lukoianova, T., & Rubin, V. (2014). Veracity roadmap: Is big data objective, truthful and credible? *Advances in Classification Research Online, 24*, 4-15.

Lyons, D. (2014). Surveillance, snowden, and big data: Capacities, consequences, critique. *Big Data & Society, 1*, 1-13.

Madden, S. (2012). From databases to big data. *IEEE Internet Computing, 16*, 4-6.

Maji, P. K., Roy, A. R., & Biswas, R. (2002). An application of soft sets in a decision making problem. *Computers & Mathematics with Applications*, 44, 1077-1083.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt.

McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review, 90*, 60-68.

Meyer, R. (2014). Everything We Know About Facebook's Secret Mood Manipulation Experiment. Atlantic. Available from https://www.theatlantic.com/technology/archive/2014/06/everything-we-know-about-facebooks-secret-mood-manipulation-experiment/373648/

Moerel, L., & Prins, C. (2016). Privacy for the Homo Digitalis: Proposal for a New Regulatory Framework for Data Protection in the Light of Big Data and the Internet of Things. Available from https://ssrn.com/abstract=2784123

Moore, A. (2008). Defining privacy. *Journal of Social Philosophy, 39*, 411-428.

Moreau, D. H. S. (2013). *Philosophical Foundations of Discrimination Law*. Oxford: Oxford University Press.

Mundie, C. (2014). Privacy pragmatism: Focus on data use, not data collection. *Foreign Affairs, 93*, 28-34.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge: MIT Press.

Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification’. *The Journal of Strategic Information Systems, 24*(1), 3-14.

OECD. (2016a). *Advanced Analytics for Better Tax Administration: Putting Data to Work*. Paris: OECD Publishing. Available from www.oecd.org/publications/advanced-analytics-for-better-tax-administration-9789264256453-en.htm.

OECD. (2016b). *Technologies for Better Tax Administration: A Practical Guide for Revenue Bodies.* Paris: OECD Publishing. Available from www.oecd.org/publications/technologies-for-better-tax-administration-9789264256439-en.htm.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge: Harvard University Press.

Payton, T. M., & Claypoole, T. (2014). *Privacy in the Age of Big Data: Recognizing Threats, Defending your Rights, and Protecting Your Family*. Lanham: Rowman & Littlefield.

Puschmann, C., & Burgess, J. (2014). Metaphors of Big Data. *International Journal of Communication, 8*, 1690-1709.

Qiang, W., Zheng, X., & Hsu, C.-H. (2015). Cloud Computing and Big Data. Second International Conference, Cham, Springer.

Regan, P. M. (1995). *Legislating Privacy: Technology, Social Values, and Public Policy*. Chapel Hill: University of North Carolina Press.

Regulation 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).

Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics and repealing Regulation (EC, Euratom) No 1101/2008 of the European Parliament and of the Council on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities, Council Regulation (EC) No 322/97 on Community Statistics, and Council Decision 89/382/EEC, Euratom establishing a Committee on the Statistical Programmes of the European Communities (Text with relevance for the EEA and for Switzerland).

Richards, N. M., & King, J. H. (2013). Three paradoxes of big data. *Stanford Law Review Online, 66*, 41-46.

Richards, N. M., & King, J. H. (2014). Big data ethics. *Wake Forest Law Review, 49,* 393-432.

Rubenstein, I. (2013). Big data: The end of privacy or a new beginning*? International Data Privacy Law, 3*, 74-87.

Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Upper Saddle River: Prentice Hall.

Satija, A., & Hu, F. B. (2014). Big Data and systematic reviews in nutritional epidemiology. *Nutrition Reviews, 72*, 737-740.

Saunders, J., Hunt, P., & Hollywood, J. S. (2016). Predictions put into practice: A Quasi-experimental evaluation of Chicago's predictive policing pilot. *Journal of Experimental Criminology, 12*, 347-371.

Schneier, B. (2016). *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World*. New York: Norton & Company.

Shapiro, M. (2005). Smart Cities: Quality of Life, Productivity, and the Growth Effects of Human Capital. National Bureau of Economic Research Working Paper 11615.

Squires, G. D. (2003). Racial profiling, insurance style: Insurance redlining and the uneven development of metropolitan areas. *Journal of Urban Affairs, 25,* 391-410.

Stevenson, D., & Wagoner, N. J. (2014). Bargaining in the Shadow of Big Data. *Florida Law Review, 67*, 1337-1399.

Tene, O., & Polonetsky, J. (2012). Privacy in the age of big data: A time for big decisions. *Stanford Law Review, 64*, 63-69.

Tene, O., & Polonetsky, J. (2013). Big data for all: Privacy and user control in the age of analytics. *Northwestern Journal of Technology and Intellectual Property, 11*, 239-272.

Tene, O., & Polonetsky, J. (2017). Taming the Golem: Challenges of Ethical Algorithmic Decision Making. Working Paper. Available from https://fpf.org/wp-content/uploads/2016/05/Golem_May153-1.docx

Thomas, E. (2016). Why Oakland Police Turned Down Predictive Policing. Available from https://motherboard.vice.com/en_us/article/ezp8zp/minority-retort-why-oakland-police-turned-down-predictive-policing

Toh, S., & Platt, R. (2013). Big data in epidemiology: Too big to fail? *Epidemiology, 24*, 938-939.

Torgler, B. (2003). Tax morale, rule-governed behaviour and trust. *Constitutional Political Economy*, *14*, 119-140.

United Nations. (2013). Resolution adopted by the Economic and Social Council on 24 July 2013 [on the recommendation of the Statistical Commission (E/2013/24)] 2013/21. Fundamental Principles of Official Statistics. Available from https://unstats.un.org/unsd/dnss/gp/FP-Rev2013-E.pdf

United Nations, Resolution adopted by the General Assembly on 29 January 2014 [without reference to a Main Committee (A/68/L.36 and Add.1)] 68/261. Fundamental Principles of Official Statistics.

van der Sloot, B. (2011). Public sector information & data protection: A plea for personal privacy settings for the re-use of PSI’. *Informatica e Diritto, 1*, 219-238.

van der Sloot, B. (2013). From data minimization to data minimummization. In B. Custers, T. Calders, B. Schermer, & T. Zarsky (Eds.). *Discrimination and Privacy in the Information Society* (pp. 273-287). Heidelberg: Springer.

van der Sloot, B., & van Schendel, S. (2016). International and Comparative Legal Study on Big Data. WRR-rapport, Working Paper 20.

van der Sloot, B. (2017). *Privacy as Virtue*. Cambridge: Intersentia.

van der Sloot, B. (2018). Elementaire deeltjes: Big Data. Amsterdam: Amsterdam University Press.

Wetenschappelijke Raad voor Regeringsbeleid. (2016). Big Data in een vrije en veilige samenleving. WRR-rapport. Amsterdam: Amsterdam University Press.

Witkowski, K. (2017). Internet of things, big data, industry 4.0 – Innovative solutions in logistics and supply chains management. *Procedia Engineering, 182*, 763-769.

Young, S. D. (2015). A "big data" approach to HIV epidemiology and prevention. *Preventive Medicine, 70*, 17-18.