Project acronym:	BYTE
Project title:	Big data roadmap and cross-disciplinarY community for addressing socieTal Externalities
Grant number:	619551
Programme:	Seventh Framework Programme for ICT
Objective:	ICT-2013.4.2 Scalable data analytics
Contract type:	Co-ordination and Support Action
Start date of project:	01 March 2014
Duration:	36 months
Website:	www.byte-project.eu

Deliverable D2.1:
Report on legal, economic, social, ethical and political issues

Author(s):	Anna Donovan, Rachel Finn and Kush Wadhwa (editor), Trilateral Research & Consulting Sertac Oruc, Claudia Werker and Scott W. Cunningham, TU Delft Guillermo vega Gorgojo, Ahmet Soylu, Dumitru Roman and Rajendra Akerkar, University of Oslo Jose Maria Garcia, UIBK Hans Lammertant, Antonella Galetta and Paul De Hert, Vrije Universiteit Brussel Stephane Grumbach and Aurelien Faravelon, Inria Alejandro Ramirez, Siemens

Dissemination level:	Public
Deliverable type:	Final
Version:	1
Submission date:	30 September 2014

Table of Contents

EXECUTIVE SUMMARY

Anna Donovan and Rachel Finn
Trilateral Research & Consulting

In this deliverable, we utilise a combination of literature review and desk research to identify a number of economic, legal, social and ethical, and political issues relevant to big data, to identify examples that illuminate these issues and display the potential positive and negative externalities that flow from these issues. We also examine where potential action can be taken to encourage the benefits that flow from some of these issues, as well identifying action that is required to minimise any negative effects of these issues. This work was undertaken as part of the EU FP7-funded project "Big data roadmap and cross disciplinary community for addressing societal externalities" (BYTE), within work Package 2 (WP2), Economic, legal, social, ethical and political issues of big data.

This document recognises a number of economic, legal, social, ethical and political issues that arise in relation to big data that can assist big data actors to unlock future potentials whilst understanding the negative implications raised by these issues. In that context, economic issues include innovation through new business models that focus on deriving added value and capturing efficiencies of big data. However, these issues are related to emerging privacy concerns. Legal issues that arise in relation to big data include intellectual property rights, licensing and contract issues, as well as data protection and privacy risks, jurisdictional issues and implications for due process. Further, big data practices such as transparency, profiling and tracking, re-use and unintended secondary use of data, open access, and levels of data access raise social and ethical issues such as trust, discrimination, inequality of access, privacy, exploitation and manipulation. Lastly, political issues emerge as a result of the challenges big data presents to relationships between states, citizens and corporations.

Specifically, our discussion in chapter two is focussed on the economic issues of big data in light of the current and potential value of big data that can boost the economy by increasing efficiency and supporting new business models and innovation. Economic issues such as innovation, entrepreneurship and management efficiency are linked to a change in business models in a variety of sectors, including retail, manufacturing, health care, life sciences and the public sector. We also consider the role of big data warehouses and data markets. It is apparent from this analysis that the challenge for big data economic actors is to realise economic and social benefits while effectively minimising the potential negative effects (e.g., privacy infringements). Each of these sectors are impacted by big data differently, in respect of the value that they can derive from the processing of large data sets. Thus, big data is a promise and challenge at the same time. While it promises huge productivity, that productivity comes with numerous issues. We examine potential solutions to the challenges presented in this context. For example, privacy issues are of great concern to private parties, and thus, economic and technological solutions must also address this area of concern. Overall, this chapter finds that big data is a new type of asset to both the private and the public sector, particularly as it represents a key basis for competition. This means that all stakeholders, particularly governments and businesses, must equip themselves with big data

and analytics to better exploit the potential of big data. In fact, big data might make the difference to how countries and companies compete and thrive.

In chapter three we discuss the legal issues of big data including intellectual property rights, licensing and contract issues, as well as data protection and privacy, jurisdictional issues and the effect that big data might have on due process. We consider these issues in terms of whether the current legal frameworks adequately address the gap between the law and technological capability. Under intellectual property laws, legal changes like introducing a new exception for data mining, enlarging the exception for technical processes or making copyright law more technology-neutral through fair use-clauses can help shorten this learning curve. On the other hand, the fundamental problem of the outmoded conceptualisation present in copyright law could remain and pose problems with other technological developments as well. Further, intellectual property rights provide a default property regime regulating access to and control of data. This default regime only establishes a starting point that can be deviated from by using contractual agreements and specific licenses. We also consider whether data placed on social networking sites, for example, are owned by the person who produced it or the person who provides the platform that collects, stores and processes it. We then identify issues raised during the application of the data protection framework to big data processing, which are not adequately addressed by the current framework. Such issues include anonymisation, consent and issues with definitions provided by the Data Protection Directive (95/46/EC). We then look at how big data processing impacts the legal doctrine of due process. We reveal that due process is more than transparency and concerns the actual capacity of a person being heard (and not just observed), to question decisions and to object to the use of certain criteria, models or data. However, transparency is a key building block in due process. Finally, we consider the relationship between accountability and liability. Accountability involves issues of liability (who is responsible for which fault), and issues of jurisdiction (which laws apply and which courts have jurisdiction over the matter). These aspects of liability are complicated by big data processing and cloud computing that remove jurisdictional boundaries. Again we reveal that the legal frameworks were conceptualised in another technological environment, which makes them less relevant to big data practices when attempting to apply traditional assessments to determine liability and jurisdiction. We conclude that the legal framework produces uncertain outcomes for economic competitiveness, particularly as the legal environment remains complex and supports overly protective regulation. This is because existing legal frameworks provide impediments to big data processing and that big data processing challenges the functioning of several of these legal frameworks. A clear need for legal reform is present in order to reap the advantages of big data, whilst protecting values that are endangered by technological developments.

In chapter four we examine the social and ethical issues that arise in relation to big data practices, as well as the negative and positive implications these issues have for individuals and society. The big data practices that implicate social and ethical issues discussed in this report are transparency, profiling and tracking, personalisation techniques, re-use, unintended secondary use, sharing, data access and open data. These practices raise social and ethical issues such as trust, discrimination, inequality of access, privacy, exploitation and consumer manipulation. For example, transparent practices can produce positive and negative implications for big data companies and users. On one hand, transparency builds user trust and in turn, promotes the disclosure of more data by trusting data subjects, whilst on the other hand, it can cause users to modify or distort their behaviour and limit the amount of data they provide or perform data sabotage. Information technology practices such as profiling and tracking can lead to a form of digital discrimination. Such discrimination requires effective

minimisation to limit the socio-economic repercussions for those discriminated against. In addition, profiling and tracking using big data can also exploit users when commercial gain is had at the expense of the social and ethical values of individuals. However, awareness of the negative externalities of big data practices such as exploitation can translate into positive outcomes for users who may motivate big data actors implementing policies that preserve ethical and social values. This is undeniable benefit for society that can also support a sustainable big data industry. Re-use and unintended secondary uses of big data can have social consequences and also raise ethical questions. The risk of this occurring is increased when those using or re-using large data sets cannot be certain of the data quality or accuracy. Big data technologies and practices that are either not universally accessible or that enable or restrict access to large data sets raises social issues relating to potential inequality of access to data. However, there may be some circumstances that warrant reduced access to data sets. For example, when the technical nature of the practices being implemented or the complexity and size of the data require expertise that is not held by all big data actors. Lastly, the availability of large data sets to the public, either through open government data or commercial open data policies and initiatives raise significant privacy issues. Recognition of negative externalities that can flow from the implementation of technologies and practices that compromise ethical values such as privacy provide a warning for big data companies operating into the future. This chapter reveals the importance of recognising that big data technologies can be used in a socially and ethically responsible manner. This can be achieved by big data companies undertaking a moral assessment of their practices and technologies in order to create a big data industry that is sustainable and influences society in a positive way.

Lastly, our discussion in chapter five focuses on the political issues that arise as a result of big data. Big data will impact politics at all levels, including: international relations between states; national governance and political institutions; and regional organisation and administration. This is because big data and the digital environment will change the balance between citizens, states and corporations. Corporations are currently challenging past equilibrium with states on many issues in a very broad spectrum of activity ranging from taxes to data protection, from utilities to copyright, for example. Corporations are challenging states, including and particularly in remote countries in a way that creates new tensions, not seen with multinationals in other areas in the past. We also consider the relationship between the state and citizens, as that relationship will be strongly impacted by big data. This is a result of two main happenings: first, big data reshapes the public space and more precisely what citizens can access and what states share; and second, big data opens a wide range of new possibilities in terms of governance. Our examination reveals the absence of corporations in the big data sector in Europe, and the increasing dependency upon US systems for services that can now be considered as utilities, restricts the capacity of Europe to react to legal disputes related to values Europe is committed to preserving. For example, the data that are handled by US corporations fall under US laws, thus leading to some new territoriality of the American laws, despite the fact that the data originates from European citizens and may be stored on installations on the European soil. The US is not the only country to store their data on their own systems, China, Russia, as well as other (mostly Asian) countries have developed strong local systems, which harvest most of the data of their population. Europe is therefore in a position of weakness that exacerbates strong negative externalities of the big data industry, which are due to the very weak political capacity on its own territory. Moreover, there is a risk that such disputes may be framed as trade disputes rather than conflicts of law, and may be vulnerable to the exercise of economic and political power through threats to cut vital services.

This report has revealed that big data per se, and the practices and technologies associated with large data sets, implicates a number of economic, legal, social, ethical and political issues, a number of which can produce positive and negative societal externalities. This report presents a preliminary investigation of some of the economic, legal, social, ethical and political issues that may be relevant to the externalities produced by big data to assist big data actors in capturing the benefits of big data whilst minimising the negative implications that big data can have on society. BYTE will use this information as a springboard to further investigate social issues in relation to specific big data case studies.

1 INTRODUCTION

Anna Donovan
Trilateral Research & Consulting

1.1 OVERVIEW

The collection, use, sharing and linking of big data implicates a number of economic, legal, social, ethical and political issues, including those which may result in positive and negative societal externalities. This report presents a preliminary investigation of some of the economic, legal, social, ethical and political issues that may be relevant to the externalities produced by big data. A variety of these issues arise in relation to big data usage across a number of participating sectors. Thus, identifying such issues can assist in a better understanding of areas for potential growth and development within the big data industry and assist in bolstering Europe's digital economy. This examination also identifies a number of negative societal externalities that arise in relation to the issues addressed which is pertinent to beginning to address these negative consequences so that they do not overwhelm the potential economic and social benefits of big data.

1.2 ECONOMIC ISSUES

Big data implicates economic issues that have positive and negative social consequences. Big data relates to the economy because it can be a catalyst for innovation, in particularly when new business models require development to incorporate strategies for deriving the added value from big data and in order to capture the efficiencies of big data across a number of sectors. However, concerns for privacy are raised along side these positive effects.

The explosion of the big data analytics sector has lead policy-makers, industry actors, journalists and academics to view data as a "resource" which has "value" that can be exploited and which can provide a boost to the economy. A review of relevant literature reveals nine major arguments about the value of big data, which can be grouped to introduce three value propositions: empirical evidence, automation and control, and information. These value propositions are similar to, but short of, business models or business cases. They are useful as propositions because they provide a lens with which to examine the phenomena, and to test various claims and counter-claims regarding the economic impact of big data, and potentially assess the big data driven boost to the economy. This boost to the economy primarily comes in the form of increasing efficiency and supporting innovative business models. In terms of efficiency, big data can assist industry and other stakeholders in making "proactive knowledge-driven business decisions"[1] and the "extraction of embedded intelligence and data insights". Other business efficiency benefits include customer intelligence, supply chain management, quality management, risk management, performance management and fraud detection. Sector-specific efficiency benefits include healthcare efficiencies, reduced strain on infrastructure, better provision of energy, greater accuracy in prediction and measurement of weather events, as well as others. Improvements in efficiency might also support innovative business models by reducing entry barriers and making it less

risky to launch new products, services or companies because of improved information and reduced uncertainty. Finally, consumers themselves may experience economic impacts through the provision of services at no cost, based on the value of the usage data generated by the service for the company. All of these discussions promise significant positive economic externalities in relation to big data. However, in terms of negative economic externalities, Manovich notes that it is primarily well-established, large companies and institutions that have access to big data sets, and which may not reduce entry barriers as significantly as argued above.

We examine how innovation and entrepreneurship are related to big data as well as business models and efficiency by considering the following questions: What are the potential economic impacts of big data? How does big data create new opportunities and market? How does it change business, business relationships and the economic landscape? How can firms, consumers, policy makers and other economic stakeholders use big data in responsible ways, e.g. by guarding individual privacy? In the following, we will discuss economic issues of big data in three ways: One, we develop a number of value propositions of big data indicating negative and positive effects emerging from it (Section 1.2). Two, we look into the potential value and hurdles of big data in different sectors (Section 1.3). Three, we discuss how big data affects innovation and entrepreneurship, management efficiency, business models as well as data markets and warehouses (Section 1.4). The legal, political, social and ethical as well as the political issues will be discussed in Section 3 to 5.

Overall, the report on economic issues associated with big data supports the view that big data is expected to be a new type of asset to both the private and the public sector as it represents a key basis for competition. All stakeholders are challenged to equip themselves with big data and analytics in order to be able exploit the potential of big data. This does not only hold for companies but also for governments. In fact, big data might make the difference in how countries and companies compete and thrive. Analytics of big data might offer solutions to problems of a struggling global economy.

1.3 LEGAL ISSUES

The collection and processing of big data also raises a number of legal issues in relation to intellectual property rights, licensing and contracting. In addition, big data has implications for privacy and data protection, as well as due process, liability and, it can lead to jurisdictional problems. The legal issues addressed in this report raise a number of tensions where the legal frameworks do not adequately or relatively address big data processing.

Potential legal issues that could raise positive or negative externalities include intellectual property rights, such as copyright, and the relationship between sui generis protection for databases as part of big data processing. These issues are also related to licensing and contracting, methods used to derogate from the copyright framework. Specifically, SME's may be disadvantaged by "a complicated legal environment and overly protective legislation"; however, it also recognises that clear market conditions may be conducive to growth, investment and innovation. An example of when complexity or ill-fitting regulation arises when attempting to apply the current data protection legal framework to big data processing. Failures to protect the security of personal data could result in data being shared with unauthorised users, particularly as data warehouses necessitate the consideration of more complex data security measures. However, in order to prevent this, the principles of the framework require attention to better apply to big data processing. Furthermore, while some

actors attempt to meet privacy concerns by anonymising data, the linking of data sets may result in the ability to re-identify individuals once disparate data sources are linked together. Big data processing also raises issues other issues concerning liability due process and jurisdictional issues. Due process is implicated because big data processing can be used to inform decisions about people or even as part of automated decision-making. Possible application areas are in marketing and targeted advertising, insurance, credit lending and even security-related activities. This opens a wide area of problems, which are partially dealt with by the data protection framework but also raise issues covered by non-discrimination law, consumer protection, etc. Big data processing activities also involve issues of liability namely, who is responsible for which fault, and issues of jurisdiction, which laws apply and which courts can deal with the problem. All these aspects of accountability become more complicated with big data processing and cloud computing as underlying infrastructure. Again we notice that the legal frameworks were conceptualized in another technological environment and with other use cases as reference. In this section we consider liability, followed by a discussion on issues of jurisdiction. Thus, this chapter reviews some of the legal frameworks that are applicable to big data processing, and highlights the "gap" between technological capability and the legal framework has uncertain outcomes for economic competitiveness.

In general stringent legal regulation can either support or restrict the development of the big data industry and require recognition in order to understand how the current framework applies to big data, as well as identifying some of the gaps in that framework.

1.4 SOCIAL AND ETHICAL ISSUES

A number of social and ethical issues arise in relation to big data practices. Big data practices such as transparency, profiling and tracking, personalisation techniques, re-use, unintended secondary use, sharing, open data and open access implicate a number of social and ethical issues including discrimination, trust, privacy, inequality of access, exploitation and manipulation. This is because big data practices deal with data from people, and this human element reflects individual social and moral codes. These issues require recognition so that big data companies and organisations can incorporate fundamental social and ethical values into big data practices and policies. Ultimately, socially and ethically responsible big data practices can support the sustainability of a European big data industry. If big data practices compromise social and ethical values then data subjects may be reluctant to provide their data, or only to the extent that it gains them access to a service. This can limit the potential growth of big data.

Specifically, this report examines transparent practices that produce positive and negative implications for big data companies and users. On one hand, transparency builds user trust and in turn, promotes the disclosure of more data by trusting data subjects, whilst on the other hand, it can cause users to modify or distort their behaviour and limit the amount of data they provide or perform data sabotage. Ultimately, transparency is the key to building user trust, which in turn, leads to a greater amount of available data. Information technology practices such as profiling and tracking can lead to a form of digital discrimination. Such discrimination requires effective minimisation to limit the socio-economic repercussions for those discriminated against. The use of personalisation techniques such as personalised and targeted advertising also raise issues of trust, exploitation and manipulation. Big data companies can build trust through providing users with personalised advertising and a more tailored online experience by predicting user preferences and conveniently providing relevant

information. If users are receptive to personalised and targeted advertising then they may feel more trusting of the websites that "know them". Alternatively, users can feel exploited by this overt form of commercialisation. In addition, profiling and tracking using big data can also exploit users when commercial gain is had at the expense of the social and ethical values of individuals. Re-use, unintended secondary use and sharing of big data can also lead to social consequences and also raise ethical questions. The risk of this occurring is increased when those using or re-using large data sets cannot be certain of the data quality or accuracy. Big data technologies and practices that are either not universally accessible or that enable or restrict access to large data sets raises social issues relating to potential inequality of access to data. However, there may be some circumstances that warrant reduced access to data sets such as when the technical nature of the practices being implemented or the complexity and size of the data require expertise that is not held by all big data actors. Finally, the availability of large data sets to the public, either through open government data or commercial open data policies and initiatives raises the issue of privacy. Open access can be, by its very nature, nature privacy invasive. Big data practices such as open data, and the mining of that data, highlight the potential compromise of ethical values such as privacy.

Therefore, recognition of negative externalities that can flow from the implementation of technologies and practices that compromise ethical values such as privacy provide a warning for big data companies operating into the future.

1.5 POLITICAL ISSUES

Lastly, a number of political issues arise in relation to big data. Big data will impact politics at all levels, namely: international relations between states; national governance and political institutions; and regional organisation and administration. All three levels are considered in this report on political issues. An analysis of all levels is important because it identifies how the digital environment will change the balance between citizens, states and corporations. For example, corporations are currently challenging past equilibrium with states on many issues in a very broad spectrum of activity ranging from taxes to data protection, from utilities to copyright. This challenge, particularly when it occurs in remote countries, creates new tensions, not seen with multinationals in other areas in the past.

Europe is primarily characterised by marginal web sites that do not harvest large amounts of personal data, and many European countries are reliant on services provided in foreign countries such as the US, many of which are becoming as necessary for the economy as utilities such as transportation or energy. Such dependency already generates tensions between regions that might have strong political consequences, and the current tensions between European data protection regulators and Google, Facebook and US cloud services are clear examples. If the capacity to harvest personal data is strategic at the level of corporations, interestingly, it is also of the uttermost importance at the level of nations. The data that are handled by US corporations for instance fall under US laws, thus leading to some new territoriality of the American laws, despite the fact that the data originates from European citizens and may be stored on installations on the European soil. Furthermore, China, Russia as well as other (mostly Asian) countries have developed strong local systems, which harvest most of the data of their population, again placing Europe at a disadvantage. The absence of corporations in the big data sector in Europe, and the increasing dependency upon US systems for services that can now be considered as utilities, restricts the capacity of Europe to react to legal disputes related to values Europe is committed to preserving. However, there is a risk that such disputes may be framed as trade disputes rather than conflicts of law, and may be

vulnerable to the exercise of economic and political power through threats to cut vital services. Thus, this chapter also integrates a geopolitical perspective to understand the political challenges facing the big data industry worldwide.

Further, this section aims to explore the nature of diplomacy and trade relationships in relation to big data and the types of externalities that are produced by these issues. It will examine the power imbalances, shifts and political pressure points that influence the flows of data as well as the demand for data across borders. Specifically, it will examine the factors that enable large companies, organisations and countries to demand or extract data, and to process it in a way that benefits their interests, possibly to the detriment of other actors. It will also identify the types of data these powerful actors gain access to and the political value that is extracted from this big data. This will result in a better understanding of the societal impact of these political forces and how they influence the possibilities for the realisation of big data innovations in Europe.

Overall, this section reveals that the absence of corporations in the big data sector in Europe, and the increasing dependency upon US systems for services that can now be considered as utilities, restricts the capacity of Europe to react to legal disputes related to values Europe is committed to preserving.

1.6 METHODOLOGY

This deliverable used mainly desk based research to provide a broad brush examination of the positive and negative externalities of each issue, the catalysts for the externalities and those affected by the externalities. This desk research was reliant on academic journal articles, project reports, media materials, materials from industry and any other relevant information. The information provided by this report will also feed into the rest of the BYTE project and represents a firm platform from which to consider the potential externalities raised in the BYTE case studies, including: environmental data, commercial data, utilities/ smart cities data, cultural data, energy data, health data, and transport data.

2 ECONOMIC ISSUES IN BIG DATA

Sertac Oruc and Claudia Werker and Scott W. Cunningham
TU Delft
With contribution from Guillermo Vega Gorgojo, Ahmet Soylu, Dumitru Roman, Rajendra Akerkar (UIO) and Jose M. Garcia (UIBK)

2.1 INTRODUCTION

Big data encompasses all walks of life, including private activities, sciences and industries.[2]

> Making the world more intelligent, identify patterns undetected before, take decisions not based on the limited experts' knowledge but on the huge mass of data from the inscrutable reality. This is the promise of big data.[3]

While big data promise solutions to problems otherwise unsolvable big data also threaten people's privacy[4]: To give an example, big data can contribute to real-time management of evacuation scenarios by using GPS localization information of mobile phones. The challenge is, however, to this in way respecting privacy , as this is a high good, as is confidentiality[5]. An example of lack of respect for privacy a supermarket chain in the United States (U.S.) detecting pregnant women based on their buying behaviour and sending specific advertisement to them[6]. The furious father of a teenage girl stormed one of the markets for sending pregnancy related advertisement to his daughter. He claimed that his daughter was not pregnant but she turned out to be pregnant after all. Interestingly, while, in general, pregnant women think that they undergo a very specific pregnancy experience they behave very similarly when buying their groceries. These and similar analyses replace individuality by hyper-individuality which is of great interest to industry and science. For example, providing tailor-made advertisements to Internet-users, as does the firm Bluekai[7], might be only first steps of developing new markets – potentially at the expense of individual privacy.

The material in this work package materially builds upon prior work of the BYTE project. Specifically we start with the definitions and big data initiatives detailed in WP 1. Starting from this work on definitions, we can offer a more focused definition of big data and economics. Economically, big data means

> economically extract(ing) value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis. There are three main characteristics of big data: the data itself, the analytics of the data, and the

[2] DNV GL AS, "DNV GL STRATEGIC RESEARCH & INNOVATION POSITION PAPER 4-2014", 27.07.2014., Helbing, Dirk, and Stefano Balietti, "From social data mining to forecasting socio-economic crises", *The European Physical Journal-Special Topics,* 195, 1, 3-68 and Smolan, Rick, and Christoph Kucklick, "Der vermessene Mensch", *GEO,* 08, August 2013, 80-98.
[3] Ibid. , p.85
[4] McKinsey & Company, "Big data: The next frontier for innovation, competition and productivity", 2011, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.
[5] Helbing, Dirk, and Stefano Balietti, "From social data mining to forecasting socio-economic crises", *The European Physical Journal-Special Topics,* 195, 1, 3-68.
[6] Smolan, Rick, and Christoph Kucklick, "Der vermessene Mensch", *GEO,* 08, August 2013, 80-98.
[7] Oracle, "Bluekai: Data Activation System Solutions for Marketers, Publishers, Agencies and Data Providers", http://bluekai.com/.

> presentation of the results of the analytics. Then there are the products and services that can be wrapped around one or all of these big data elements. [8]

So, following Gantz and Reinsel[9] we go way beyond looking into the effects of the three main technical characteristics of big data, i.e. volume, velocity and variety. We investigate the socioeconomic value potentially emerging from the use of big data as a new factor of production[10]. This value is the ultimate incentive for stakeholders, such as firms, governments, and consumers, to engage in accumulating, processing and using big data. Big data differs from traditional data because its amount is so huge that it cannot be collected, stored, shared and analysed by traditional data analysis but requires new strategies and algorithms[11]. Terabytes of data about consumers, firms, production and marketization processes etc. form big data (c.f. this and the following[12]). In particular, smart phones, smart meters, surveillance cameras, automobiles as well as ever increasing Internet activity of consumers, firms, scientist and policy makers contribute to big data. In principle, all stakeholders have the possibility to create, collect, store, share, process and analyse big data. The amount of data, which is referred as "digital universe", is expanding ever since the origin of the Internet with a rate of more than twice per year. The majority of the digital universe is created and consumed by the consumers by watching digital TV, interacting on social media and using the Internet. 80% of all these data is under the liability and responsibility of the enterprises, such as Facebook, Google and the like. However not all of these data is economically meaningful for big data analytics. While the amount of data being created is enormous so far its potential is only used to a very limited extent[13].

In the remainder of Section 1 we analyse economic issues emerging from the chances and threats big data brings for society as well as the impacts big data might have on the economy and the business landscape worldwide. Economic issues indicate positive or negative economic implications of big data. We do not use the term economic externalities here as this means either the cost or the benefit affecting an economic agent, i.e. a firm or consumer, who neither choose for that cost or benefit nor being compensated for it. Here, we analyse a broader set of questions, i.e. economic issues that might as well occur because of conscious choice of economic agents. We do so in order to not unduly limit our analysis.

Thereby we will contribute to answering questions such as: What are the potential economic impacts of big data? How does big data create new opportunities and market? How does it change business, business relationships and the economic landscape? How can firms, consumers, policy makers and other economic stakeholders use big data in responsible ways, e.g. by guarding individual privacy? In the following, we will discuss economic issues of big data in three ways: One, we develop a number of value propositions of big data indicating negative and positive effects emerging from it (Section 1.2). Two, we look into the potential value and hurdles of big data in different sectors (Section 1.3). Three, we discuss how big data affects innovation and entrepreneurship, management efficiency, business models as well

[8] Gantz, John, and David Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east", *IDC iView: IDC Analyze the Future*.
[9] Ibid.
[10] Jones, Steve, *Why 'Big Data' is the fourth factor of production*, 2012, http://www.ft.com/cms/s/0/5086d700-504a-11e2-9b66-00144feab49a.html#axzz38JHElWO4.
[11] McKinsey & Company, "Big data: The next frontier for innovation, competition and productivity", 2011, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.
[12] Gantz, John, and David Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east", *IDC iView: IDC Analyze the Future*.
[13] Ibid.

as data markets and warehouses (Section 1.4). The legal, political, social and ethical as well as the political issues will be discussed in Section 3 to 5.

2.2 VALUE PROPOSITIONS OF BIG DATA

As data is considered as the fourth production factor on top of land, capital and labour[14] we give an overview of the value propositions of this fourth production factor. Factors of production are embodied in material goods, components and architectures of technology, as well as skills and routines. Given these considerations of economic productivity, how is big data embodied in the marketplace?

The premise that big data is a single technology, or even an individual assemblage of technology components, is inherently flawed. To our knowledge there exists no standardized ontology of big data technologies, nor could the existing definitions of big data permit an incisive functional decomposition of big data into individual technologies. The idea that big data represents skills, knowledge and routines, and is therefore a multiplier on labour has considerably more merit. More persuasive still is the argument that big data is a "general purpose technology" and not a specific set of organizational routines or technological functions.

2.2.1 The Economic Embedding of Big Data

General purpose technologies[15] are technologies which have all-pervasive impacts on the economy, stimulating structural change across multiple industrial sectors. The very idea that big data is a general purpose technology challenges the idea that there is a unitary source of economic value in big data. In this light, in the remainder of this section we review and present a number of different ideas about the real and potential value of big data. Another consequence of the view that big data is a general purpose technology is earlier discussions of the economic value of information technologies must be taken on board, further updated and reconsidered.

The literature on economic impacts of information is vast, and somewhat contradictory. Larger firms do seem to achieve higher returns on information technology investment. Nonetheless, larger firms also achieve higher economic returns per customer, and also require a larger routine investment in support systems. This counters the idea that there is a simple input-output relationship of more information technology leading to higher economic growth and economic performance. The most definitive accounts of organizational productivity and information technology investment argue for a more nuanced relationship between organizational capabilities and information technologies[16]. Taken together, these ideas of general purpose technologies and general organizational capabilities set forth an approach to more fully analyse the economic, and other impacts, of big data on society.

---

[14] Jones, Steve, *Why 'Big Data' is the fourth factor of production*, 2012, http://www.ft.com/cms/s/0/5086d700-504a-11e2-9b66-00144feab49a.html#axzz38JHElWO4.

[15] Breshnahan, T. F. , and M. Trajtenberg, "General purpose technologies: Engines of Growth", *Journal of Econometrics,* 65, 1, 83-108.

[16] Melville, N., K. Kraemer and V. Gurbaxani, "Review: Information technology and organizational performance: An integrative model of IT business value", *MIS Quarterly,* 28, 2, 283-322.

2.2.2 Value Propositions for Big Data

The purpose of the following section is to set forth a design for further systematic monitoring of big data activities across multiple sectors. The goal here is to clarify current arguments about who is benefiting from big data and how big data impacts others. A more complete appraisal of the impacts will be provided in year two of the BYTE project.

As it turns out there is not unanimity of opinion about the impact of big data. The goals of this section are therefore to perform a discourse analysis[17] of some of the leading, popular texts on big data. This can be used to identify opinions, clarify strategic arguments, focus monitoring activities and develop hypotheses for further testing.

A survey and scan of the literature reveals nine major arguments about the value of big data. These nine major arguments may be grouped by family to introduce three value propositions: empirical evidence, automation and control, and information. These value propositions are short of business models or business cases. They are useful as propositions nonetheless, because they provide a lens with which to examine the phenomena, and to test various claims and counter-claims regarding the economic impact of big data. Furthermore, these propositions reveal that there may be a multiplicity of reasons for adopting big data techniques. We summarize the three value propositions for big data in Table 1 and discuss them in further detail below.

Value Proposition	Themes and Variations
Empirical Knowledge about the World	The value of big data lies in its capability for improving decision-making [Empiricism]
	There is no value in big data, rather it is necessary to adopt big data merely to stay ahead of competition [Red Queen]
	The value of big data lies in enhanced trade-offs between organizational exploration and exploitation [Exploration]
Automation and Control of Socio-economic Processes	The value of big data lies in the potential for future automation [Automation]
	The value of big data lies in the capacity to improve social and technical control of systems [Control]
	The value of big data lies in enhancing the extractive potential of certain segments of society [Extraction]
Information	There is no inherent value in big data, rather it is a side-effect of enhanced inter-organizational communication [Signalling]
	The value of big data lies either in exploiting, or in removing, market imperfections [Asymmetry]
	There is no inherent value in big data, rather it is a fundamental transformation in societal infrastructure and organization [Autopoeisis]

Table 1: Value Propositions for Big Data

A primary proposition for adopting big data is that it enhances empirical evidence of the world and thereby improves decision-making. These are the principal arguments of Silver[18]

[17] Gee, J. P. (2005). *An Introduction to Discourse Analysis: Theory and Method.* London: Routledge.
[18] Silver, N., *The Signal and the Noise: The Art and Science of Prediction*, Penguin, London, 2013.

and Mayer-Schönberger and Cukier[19]. Silver in particular offers a balanced account of the respective roles of evidence, and prior theory, in creating new knowledge. On the other hand, other authors are much more strident about the pre-eminence of evidence and the obsolescence of theory science[20].

There are two variations on the empirical argument – one, that big data is necessity in a ruthless economic environment, and two, that big data enables new discoveries to be brought to light and economically exploited more rapidly. Morozov[21] describes the economic logic of big data companies. Morozov argues such companies propound naïve and technocratic solutions to complex societal problems. The influence of such companies in establishing standards means that the world as a whole must embrace these solutions, even when they are fundamentally flawed. In short, big data introduces more complexity without creating more value. Eisenhardt and Tabrizi[22] in a variant on the empirical argument, emphasize the role of new information in creating flexible, adaptable, and innovative firms.

A second proposition for adopting big data is that it is a natural outgrowth of an industrial society, because it enhances the automation and control of socio-economic processes. In its principle form this proposition argues that big data enhances the productive frontier of society. This proposition comes in two variations – one is that big data enhances social control (rather than material control), and the second is that big data enhances the extractive capabilities of the societal elite. Analysing long-term time series of economic production, Hanson[23] anticipates further highly dramatic increases in the coming three decades. The source of these productive gains is the enhanced creation and diffusion of scientific knowledge. Thus Hanson also combines elements of the empirical proposition in his claims for future growth.

The social form of the automation and control argument was presaged as early as 1958 by a Harvard economist[24]. As a further example of the social arguments, Erdem and Keane[25] describe the use of data in managing consumer goods markets. See also Lanier[26], a book which describes the automating, and therefore de-humanising, characteristics of big data. Collier[27] may best represent the extraction by elite argument. Collier's work describes the role of natural resource extraction by societal (and world) elite. It must be acknowledged that Collier's argument is not primarily about the role of data and the internet, but about the political economy of society as a whole.

A third proposition for adopting big data are a host of informational arguments. The principal claim is that big data is a commodity in the new information age. Markets are fundamentally

[19] Mayer-Schönberger, V., and K. Cukier, *Big Data: A Revolution that Will Transform How we Live, Work and Think*, Eamon Dolan/Houghton Mifflin Harcourt, Boston, MA, 2013.
[20] Anderson, C., *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, 2008.
[21] Morozov, Evgeny, *To save everything, click here: Technology, solutionism, and the urge to fix problems that don't exist*, Penguin UK, 2013.
[22] Eisenhardt, K. M., and B. N. Tabrizi, "Accelerating adaptive processes: Product innovation in the global computer industry", *Administrative Science Quarterly,* 40, 1, 84-100.
[23] Hanson, Robin, *Long-term growth as a sequence of exponential modes*, Citeseer, 1998.
[24] Galbraith, J. K., *The Affluent Society*, Houghton Mifflin, Boston, 1958.
[25] Erdem, T. , and M. P. Keane, "Decision-making under uncertainty: Capturing Dynamic processes in turbulent consumer goods markets", *Marketing Science,* 15, 1, 1-20.
[26] Lanier, J. , *You Are Not a Gadget: A Manifesto*, Alfred A. Knopf, New York, 2010.
[27] Collier, P. , *The Plundered Planet: Why We Must, and How We Can, Manage Nature for Global Prosperity*, Oxford University Press, Oxford, 2010.

imperfect, and those who can harvest information can exploit these imperfections. Lanier[28] develops an argument of this sort when he discusses how big data companies develop silos of content from creative individuals which they then market as their own intellectual capital. The informational proposition for big data comes in at least two variations. One is that big data enables the exchange of economically meaningful signals between diverse organizations[29]. These signals may be inexplicable to those outside, but highly consequential to participants given the correct context. And as a second variation there is the proposition is that big data is part of a more general process of societal transformation[30].

2.2.3 Assessing the propositions

Having assembled three major value propositions of big data, in nine variations, it is now useful to consider which one of these value propositions are most likely to be true? All the propositions have their flaws and counter-arguments. The empirical argument, at least in its extreme form, is unlikely to be true. Oldroyd[31] describes the continuing tension between evidence and theory in the long history of the scientific method. It is clear from this survey of the scientific method that a balance of both theory and evidence has always been needed, and that our era is not the first to attempt placing science exclusively on an empirical foundation. Other prior eras have also attempted to place the creation of knowledge on an exclusively empirical foundation, and have failed.

In its milder version, the empirical argument also seems unlikely. Returns to big data through improved decision-making could only be a big win for all but the lowest margin industries. On the other hand, it must be acknowledge the economic value of innovation is very high, and any activity which promotes enhanced exploration of knowledge is likely to be highly valuable[32]. Finally, the premise that the most economically valuable data is complete and readily accessible should be questioned.

The automation and control argument is, on the face of it, rather more convincing. However, this can only take place if automation and control technologies are being purchased alongside big data acquisitions. These should be complementary assets. The fact that the business model of the internet is so heavily based on advertising[33] is a further argument for the potential role of big data in social control. Perhaps most convincing of all are the various informational arguments. A general counter-argument to this proposition questions whether information, in and of itself, is of sufficient economic value to warrant the substantial investment of society.

Given the previous arguments that big data is likely a general purpose technology, we should expect that multiple value propositions are at work. Different propositions may be operating in different industries. In the following sections (1.3) we examine the various business models across various sectors. The uptake of big data by sector is itself a partial indicator of the kinds of value propositions which are at play in the economy.

---

[28] Lanier, J., *Who Owns the Future?*, Simon & Schuster, San Jose, 2013.
[29] Spence, M., "Signaling in retrospect and the informational structure of markets", *American Economic Review,* 92, 3, 434-59.
[30] Kurzweil, R., *The Singularity is Near*, Viking Press, New York, 2005.
[31] Oldroyd, D. , *The Arch of Knowledge: An Introductory Study of the History of the Philosophy and Methodology of Science*, Metheun New York, 1986.
[32] Solow, R., "Technical change and the aggregate production function", *Review of Economics and Statistics,* 39, 3, 312-20.
[33] Zuckerman, E., *The Internet's Original Sin*, 2014.

2.2.4 Positive and negative impacts of big data

The value proposition and the impacts of big data are fundamentally linked. The impacts, both positive and negative, are related to the three propositions stated. For an overview see Table 2. To give an example, if big data is primarily a proposition about the collating of empirical evidence then we may see enhances in scientific discovery and the more rapid commercialization of economically valuable goods and services. On the other hand, the pervasiveness of these technologies may entail job losses in many white-collar jobs. Furthermore the necessity of big data for competitiveness, and the need for major investment, may create barriers which limit new market entrants.

In the space of automation and control, big data technologies may enhance production, and generate new-found efficiencies. Furthermore, there may be improvements in the provision of services, both routine and crisis-related. On the other hand, these technologies of control may extend still further into the public realm, resulting in the abuses of private information and potentially new forms of totalitarian government. Social control by definition is coercive in character, and may create deadweight losses in society.

Informational impacts of big data depend in part on how the technologies intersect with the current economic marketplace. If big data technologies are affordable, ubiquitous, and readily employed, then their use may lead to new found efficiencies. Furthermore existing boundaries between organizations may collapse, leading to new found sources of flexibility and coordination. However the expanded role of exchanges could be corrosive to the public sphere, and other non-market forms of coordination in general. On the other hand, if big data technologies are employed by an elite segment of the market then rent-seeking may result. Big data will enhance the bargaining power of selected market actors. Furthermore, the market may be exposed to higher-levels of systematic risk given behind the scenes coordination. Big data may also be an epiphenomena of ever-increasing levels of complexity in economic life. Such complexity could easily become overwhelming for many economic agents of today.

Moreover, big data and the tools for analysing data could also even lead to inefficiencies. The example of high-frequency trading on commodity market shows transactions do not necessarily lead to prices that better mirror relevant information[34]. Investigations of the E-Mini S&P which is a stock market index futures contract traded on the Chicago Mercantile Exchange's Globex electronic trading platform that in October 2012 60-70% of the price changes were due to self-generated activities. In 2005 this was only about 20-30%. This means that the price changes have been increasingly caused by the automated and quick analysis tools. As a consequence prices in these markets are less informative and the markets are more inefficient.

Value Proposition	Positive Impacts	Negative Impacts
Empirical	• Enhances in scientific	• Employment losses for certain

[34] Filimonov, Vladimir, David Bicchetti, Nicolas Maystre and Didier Sornette, "Quantification of the high level of endogeneity and of structural regime shifts in commodity markets", *Journal of International Money and Finance,* 42, 174-92.

Knowledge about the World	discovery • More rapid commercialization of offerings	job categories • Barriers to market entry
Automation and Control of Socio-economic Processes	• Enhanced productive capability in society • Enhanced provision of routine and crisis services • Enhanced efficiency	• Invasive use of information • Totalitarian control • Dead-weight losses caused by advertising and consumption
Information	• Newly efficient markets • Improved flexibility and coordination	• Enhanced exposure to systematic risk • Informational rent-seeking • Dissolution of the public arena • Overwhelming complexity

Table 2: Value Proposition as well as their Positive and Negative Impacts

Furthermore big data would allow profiling populations for the need of the organizations to customize their actions. Profiling the customers for personal advertisement is one way of using profiling. This approach can be useful for public sector where traditionally all citizens are treated in the same way.

2.3 POTENTIAL VALUE AND POTENTIAL HURDLES OF BIG DATA IN DIFFERENT SECTORS

As big data encompasses all walks of life it might affect all sectors. However, the effect on sectors might differ considerably[35]. Using big data requires a broad set of skills and capabilities of three kinds, i.e. "… a culture for data and business development; the skills and knowledge to handle and analyse data and to build models using the data; and the availability of tools and infrastructure."[36]. In particular, scalability, flexibility, tools for analytics and visualization and data governance are important in this context[37].

When looking into online advertisement, health care, utilities, logistics and transport, and public administration which in 2010 together account for roughly one-quarter of total value added in OECD countries they seem to profit from big data substantially[38]. Advantages are the enhancement of data-driven R&D, the development of new products, optimization of production and delivery processes, marketing improvement by using targeted advertisements and personalised recommendations, the development of new organization and management tools.

For the U.S. Figure 1 shows the relationship between productivity growth of sectors (x-axis) and value potential of big data for them (y-axis)[39]. While productivity growth is measured straightforwardly for the years 2000-2008 the value potential is measured by the big data value potential index comprising five indicators: "(1) the amount of data available for use and

[35] OECD, *Exploring Data-Driven Innovation as a New Source of Growth*, OECD Publishing, 2013.
[36] DNV GL AS, "DNV GL STRATEGIC RESEARCH & INNOVATION POSITION PAPER 4-2014", 27.07.2014.
[37] Ibid.
[38] OECD, *Exploring Data-Driven Innovation as a New Source of Growth*, OECD Publishing, 2013.
[39] McKinsey & Company, "Big data: The next frontier for innovation, competition and productivity", 2011, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

analysis; (2) variability in performance; (3) number of stakeholders (customers and suppliers) with which an organization deals on average; (4) transaction intensity; and (5) turbulence inherent in a sector"[40]. The sectors are indicated by bubbles which indicate their relative share in GDP. They come in clusters indicating their specific chances and challenges. Not surprisingly computer and electronic products as well as the information sectors (Cluster A) have the greatest value potential from big data. This is not only due to their knowledge in this field but also rooted in their international knowledge and trade connections and their very strong productivity growth in the last years. Finance and insurance as well as government (Cluster B) will be able to gain from big data if they can overcome barriers to its use. A number of sectors including construction have undergone negative productivity growth (Cluster C). Unfortunately, they also seem to have a rather low potential of benefitting from big data. Sectors, such as manufacturing and whole-sale trade, mostly globally traded (Cluster D), have not only benefitted from relatively high productivity growth but could also make use of their relatively high value potential from big data. While local services (mainly Cluster E) have experienced lower growth rates they still have promising value potential from big data.

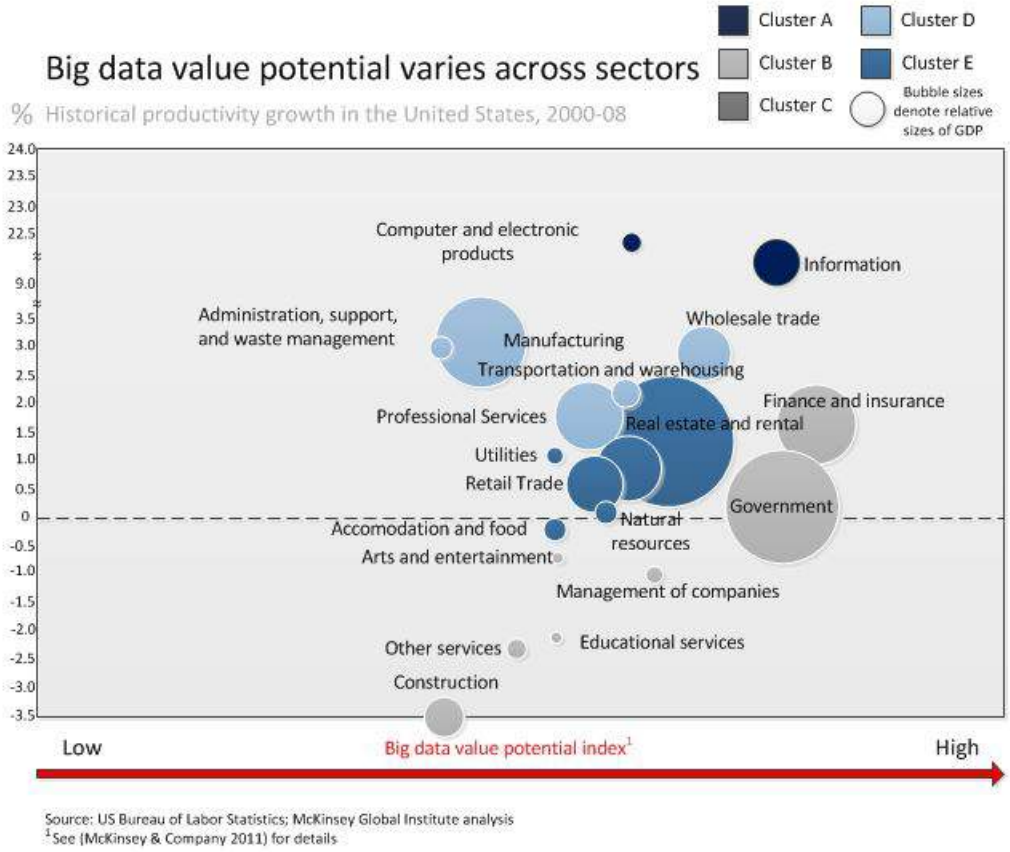


Figure 1: Productivity growth and value potential of big data in the U.S.[41]

To benefit from big data all sectors will have to deal with barriers of various kinds[42]. However, these barriers will be structurally higher for some of the sectors (see Figure 2) The

[40] Ibid., p.8
[41] Ibid., p.8
[42] Ibid.

overall ease of capture index contains four elements, all of which are shown in Figure 2: Talent relates to the relevant knowledge organizations in the sectors have in-house, IT-intensity to IT assets of a sector, data-driven mind-set to how receptive the sectors' organizations are to using big data to create value, and data availability to the proprietary corpus of data[43]. All sectors face some hurdles, such as lack of talent, lack of IT-intensity, lack of data-driven mind-set or lack of data availability. However, while sectors, such as manufacturing, are expected to face almost no hurdles the public sector – notwithstanding his high value potential – will have to develop not only a data-driven mind-set but also suitable datasets.

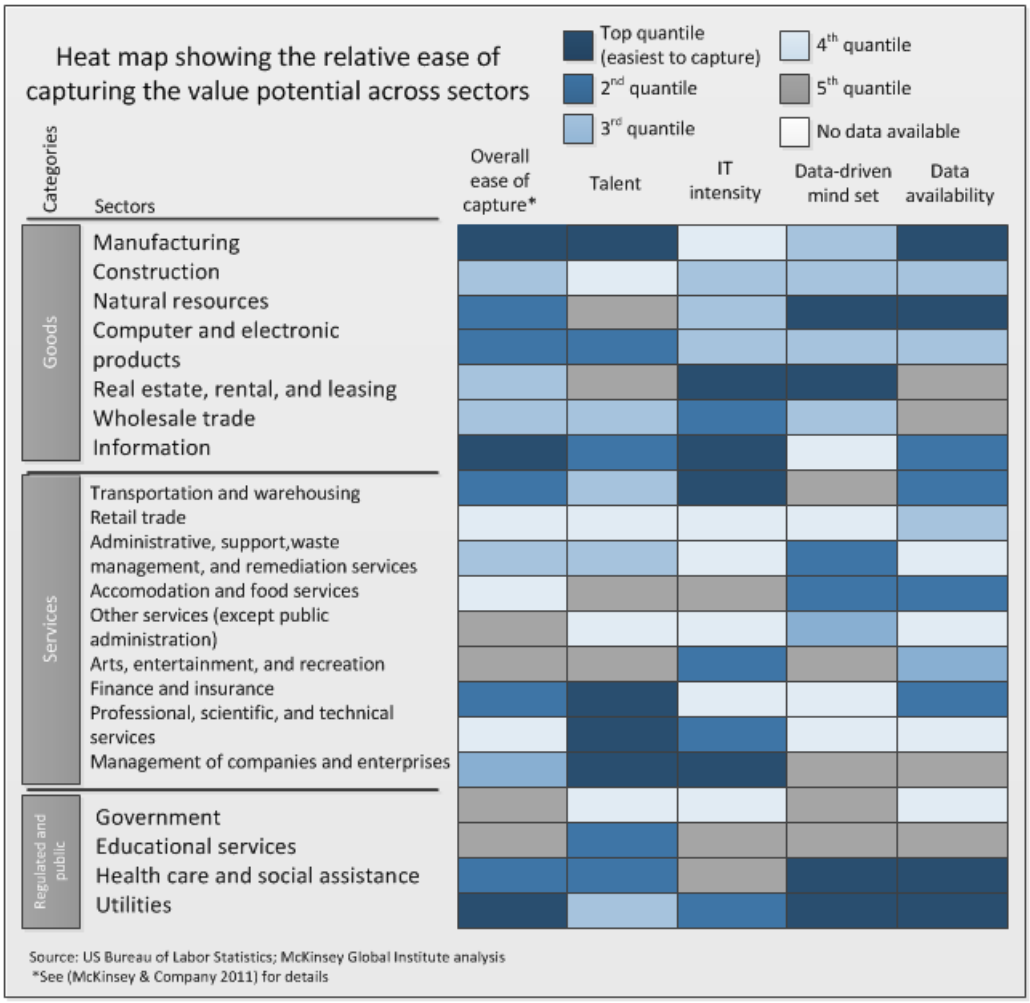


Figure 2: Ease of capturing potential value of big data for different sectors in the U.S.[44]

While big data has a huge value potential the U.S. alone is 140,000 to 190,000 people with deep analytical skills and 1.5 million managers and analysts able to analyse big data in order their decisions on the findings[45]. Moreover, there will be additional challenges in terms of investigating in appropriate infrastructure, of implementing supporting incentive systems and

[43] Ibid., pp. 124f
[44] Ibid., p.10
[45] Ibid.

laws that enables competition and innovation, of educating stakeholders about economic benefits as well as problems of big data.

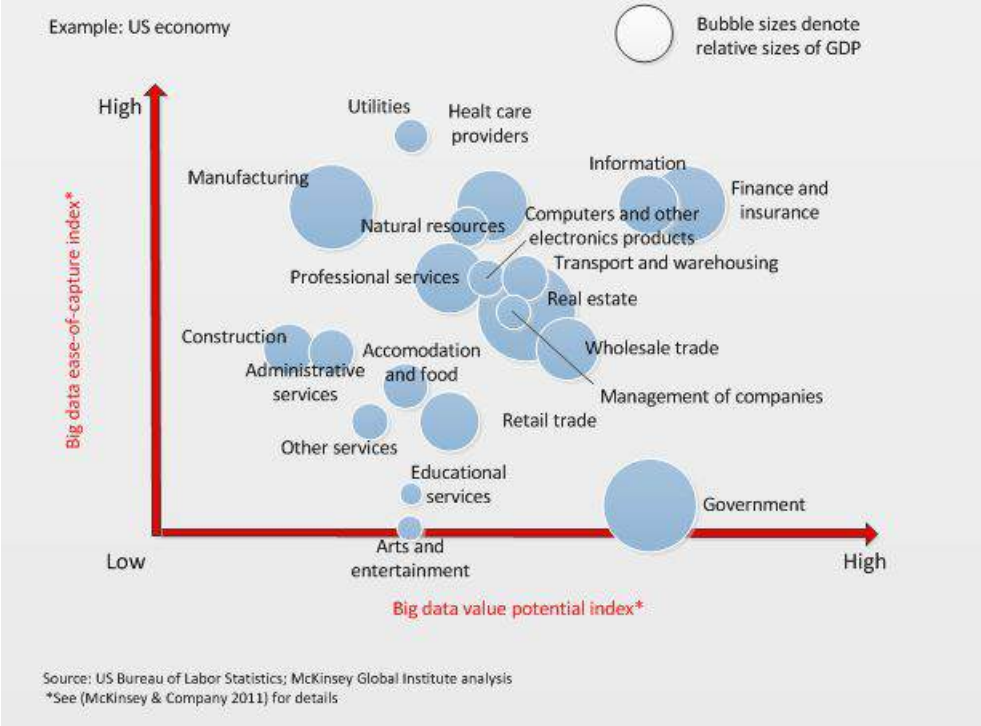


Figure 3: Big data in various sectors: value potential and ease of capture[46]

Stakeholders have completely different starting-points when it comes to big data analysis. This can be shown with the help of Figure 3. In Figure 3 we find the value potential index (x-axis) and the ease-of-capture index (y-axis) combined for various sector in the U.S. Here, we see that both finance and insurance score high on both indexes, meaning that they would not only benefit greatly from big data but that it is also easy for them to include big data analysis in their business models. While government would benefit greatly from big data analysis as well their starting point is not promising as they cannot easily adopt big data analysis. At the same time government should severely invest in big data knowledge in order to use its great big data value potential. In contrast to government the manufacturing sector has everything in house to use big data analysis. However, their value added of doing so is rather low. So, while the application of big data analysis is straightforward in finance and insurance, it might be a waste of time and resources in manufacturing..

2.4 Economic Issues of Big Data

2.4.1 Innovation, Entrepreneurship and Management Efficiency

Big data stems from all kinds of sources. The biggest source of big data analytics stem from daily transactional data processed through enterprise computers[47]. This kind of data has great

[46] Brown, Brad, Michael Chui and James Manyika, "Are you ready for the era of 'big data'", *McKinsey Quarterly,* 4, 24-35.
[47] Gantz, John, and David Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east", *IDC iView: IDC Analyze the Future*.

potential for business applications. Another vital source is data originating from entertainment and media of data about individuals. Here, trends of target groups detected from sentiments regarding different choices can be used for all kinds of ends help from marketing activities to predicting the elections. The available data of individuals on social media does comprise texts, connotations of likes or dislikes as well as images. As sophisticated tagging algorithms can detect the sentiment of individuals big data can also make use of consumer images to deduct information. Therefore, data from social media can also be applied to business purposes. The torrents of data from social media are as old as the social media itself. Facebook has existed since 2004 and Twitter since 2006. Also smart phones and other mobile devices have been only broadly used in the last decade. As the amounts and the kind of data changed substantially the query based classical databases are not always suitable for storing and processing the data. At the same time there are affordable cloud computing storage, open source software for processing large volumes of data and increasing number of big data sets available for the greater public[48]. In particular, declining costs of all elements of computing, i.e. storage, memory, processing, bandwidth, means that the previously expensive data-intensive approaches have becoming more accessible and available for all firms including small and medium sized enterprises (SMEs)[49].

As the fourth production factor big data has changed the way economic processes work[50]. While traditional data collected by firms, e.g. on their business transactions can be collected, shared and analysed by well-known technologies big data requires new technologies.

> The new data reality has created novel possibilities for business and value creation. This is the core of the big data shift; the ability to use data to obtain actionable knowledge, insights, and predictions and, eventually, to automate this process[51].

As big data has such huge value potential and as computing has become more affordable there are many new applications of big data developed SMEs. The number of examples of small companies using big data for their business are manifold. The following three examples give an idea about the various solutions, businesses and markets big data can contribute to.

- The firm Affectiva applies advanced computer vision and machine learning techniques to read tacit expressions on the faces of individuals (www.affdex.com). With the help of a cloud based emotion analysis, Affectiva can read the individuals face expressions identifying their emotional states such as surprise, dislike and attention. With this Affectiva develops solutions in marketing research for their clients.
- Another company making use of the omnipresent big data is called Bluefin Labs (www.bluefinlabs.com). Based on fifteen years research on cognitive science and machine learning at MIT Media Lab, Bluefin is able to measure viewers' engagement with television shows and ads. They have used social media commentary from Twitter, Facebook and blogs to do so.
- Yet another company is Ginger.io (http://ginger.io): It collects data about people's daily behaviour by cell phones. Ginger relies on predictive models to be applied on large amounts of data collected via a cloud based system. By collecting and analysing big data in this way Ginger is able to reveal important health patterns such as early

[48] Moss, Frank, "How Small Businesses Are Innovating With 'Big Data'".
[49] McAfee, Andrew, Erik Brynjolfsson, Thomas H Davenport, DJ Patil and Dominic Barton, "Big Data", *The management revolution. Harvard Bus Rev,* 90, 10, 61-67.
[50] Jones, Steve, *Why 'Big Data' is the fourth factor of production*, 2012, http://www.ft.com/cms/s/0/5086d700-504a-11e2-9b66-00144feab49a.html#axzz38JHElWO4.
[51] DNV GL AS, "DNV GL STRATEGIC RESEARCH & INNOVATION POSITION PAPER 4-2014", 27.07.2014.

signs of illnesses. This information allows caretakers, healthcare provides, physicians to provide better care, thereby contributing solutions for our ageing society.

In the meantime some of the formerly small companies originating in the digital universe such as Google, Facebook, Twitter and Amazon have grown in large and internationally active companies. They are already masters of big data and use it to guide decisions particularly on the executive level[52]. The potential and ability to gain a competitive edge from big data might be much bigger for large companies not stemming from the digital universe but first have to enter it. It will require their decision makers on the executive level to alter the understanding of decision making. In particular, they have to base their decision on evidence. As a consequence the organizations themselves need to refine their stance to decision making from intuition based decision making to evidence based one.

Big data analyses has the potential of making a huge difference in management efficiency. In particular, it might lead to altering existing business models substantially. Already today firms using big data in their decision making benefit from this. Based on a detailed survey of 179 large publicly traded firms collecting information on their business practices and information technology investments Brynjolfsson et al.[53] showed that firms using data and business analytics to guide decisions perform 5-6% better in terms of output and performance compared to those not doing so. Generally, it is expected that organizations that are better networked can gain competitive advantage by opening information conduits internally and by engaging customers and suppliers through Web-based exchanges of information[54]. Also according to McAfee[55], the data driven decisions are better decisions. The big data enables decision makers to decide on the basis of evidence rather than gut feeling and intuition. Even based on this fact, it has a potential to revolutionize management and decision making.

Tallon[56] reminds us that data can become costly and obsolescent. The trade-offs between the economic rent received from data, and the natural expenses incurred in maintaining and safeguarding the data, mean that data must eventually be wiped. Data, and the computer systems in which data is stored, incur organizational risks. These risks incur costs which can only partially be obviated through higher investment in safeguards. Such risks include human error, technical failure, deliberate sabotage, and systemic risks. Future innovations in big data may lower risks, and thereby costs, further enhancing the economic value of big data.

2.4.2 Efficient Functioning of Markets

Big data has the potential to transform the functioning of markets. In order to treat this claim more analytically it is important to first analyse the attributes of perfect competition, and then set forth the technologies or attributes of big data which might challenge or overturn these attributes. Before proceeding further however it is important to note that perfect competition is an idealized lens for examining economic activity.

---

[52] McAfee, Andrew, Erik Brynjolfsson, Thomas H Davenport, DJ Patil and Dominic Barton, "Big Data", *The management revolution. Harvard Bus Rev,* 90, 10, 61-67.
[53] Brynjolfsson, Erik, Lorin Hitt and Heekyung Kim, "Strength in Numbers: How does data-driven decision-making affect firm performance?".
[54] Bughin, Jacques, and Michael Chui, "The rise of the networked enterprise: Web 2.0 finds its payday", *McKinsey quarterly,* 4, 3-8.
[55] McAfee, Andrew, Erik Brynjolfsson, Thomas H Davenport, DJ Patil and Dominic Barton, "Big Data", *The management revolution. Harvard Bus Rev,* 90, 10, 61-67.
[56] Tallon, P. P. , "Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost", *Computer,* 6.

Real economic activity exists on a continuum between a perfectly competitive market, and a bureau or hierarchy which is partially or wholly shielded from the rigors of competition. Some economic researchers argue that the majority of economic activity exists in the middle of this continuum. We can therefore examine big data influences as "pushing for markets" or "pushing for hierarchies" in terms of its transformative effects on economic activity.

The assumptions of a perfect marketplace are (1) a homogeneous product, (2) many sellers, (3) perfect information, and (4) freedom of entry and exit.[57] These assumptions are perfectly standard and are common to a century of microeconomic thought. Nonetheless activities in big data significantly challenge all four of these assumptions. Let's examine each assumption in turn.

The assumption of a homogeneous product is challenged by big data. In principle big data could help purchasers source and verify the quality of their products, thus increasing the homogeneity of their purchases in terms of quality or other hard to verify attributes. Examples of this include supply chain management applications. Nonetheless the proliferation of e-commerce platforms has tended to permit a greater variety of goods to be traded than ever before. These goods are also sold across wider distances, and at an arms-length from the seller. Examples of this trend towards heterogeneity include increased sale of books, music, apps, and fashion. In terms of raw economic activity then, big data has tended to create a more heterogeneous marketplace.

Big data, by and large, has increased the number of sellers in the world, thereby increasing economic efficiency. Such claims are warranted given wider participation in general in the economic marketplace. This includes participation of small trading companies as well as in the personal market for labor. There is understandable concerns however that the big data strategies of large retailers (including Amazon and Walmart) may lead to *de facto* monopolies in many markets.

By and large big data has led to greater information perfection. Information perfection is difficult to define without substantial analytical treatment. The concept of information here signifies data which has the potential to change economic plans or behaviour. The concept of perfection indicates that all parties in an economic chain of activity have sufficient data to make appropriately contingent plans and actions. If anything the concern here is not that there is greater information perfection in the market, it is that greater information is available to some parties and not others. (This is a related concept of information asymmetry.)[58]

The fourth and final assumption of perfect competition is freedom of entry and exit. Here big data has almost certainly lessened competition. In many industries specialized computational resources and knowledge are now necessary to enter and compete. Exiting the market is also effectively much harder. Complete records of economic activity, which cannot be credibly removed from the public sphere, are widely maintained.

The competitive functioning of markets is a complex topic, and the effects of big data are nuanced. A complete analysis may find different impacts in different sectors. A more complete case study investigation of this is in fact part of the BYTE project activities for year

[57] Lipsey, R. G., Harbury, C. D. (1993), "Perfect Competition," First Principles of Economics, p. 154, Weidenfeld & Nicolson: London.
[58] Rasmusen, E. (2007), "Information," Games and Information, Fourth Edition, Blackwell Publishing, Malden: MA.

two. Nonetheless we offer the following table in hope of summarising the broadest impacts of big data on the marketplace.

Assumptions	Impacts
(1) a homogeneous product,	Big data pushes towards more hierarchies and bureaus.
(2) many sellers,	Big data pushes towards more market activities.
(3) perfect information,	Big data pushes towards more market activities
(4) freedom of entry and exit.	Big data pushes towards more hierarchies and bureaus.

Table 3. Impacts of Big Data on Economic Functioning

Big data is having a mixed role in the functioning of markets. While some trends push towards more specialized hierarchies (heterogeneous products, a lack of freedom for entry and exit), others are permitting more activities to be exposed to the market (many sellers, perfect information). Thus, big data may be fostering more complex institutional forms – neither market nor hierarchy – in the "thick middle" of day-to-day economic and commercial activities.[59]

2.4.3 Business Models

Big Data Changing Business Models

Using big data can lead to substantial changes in how production and business processes run. This can affect marketing (e.g. in terms of cross-selling or customer micro-segmentation), merchandising (e.g. pricing optimization), operations (e.g. return analysis), supply chain (e.g. inventory management) as well as new business models[60].
Big data changes the business models applied, not only in the private but also in the public sector. Following the distinction used by Stabell and Fjeldstad[61], we distinguish three business models, i.e.

- value chain
- value shop
- value network

Using these three business models it is possible to understand how big data can change the mode of operation in each type of business model[62]. For an overview with examples see Table 3. The value chain model refers to business models where the end product is a physical product. A company producing such a physical product assembles processes and production factors in order to transform raw material into the end product. Examples are the manufacturing and retailing sector (see Section 1.4.3.2). The value shop model refers to a business model designed to solve customers' problems. Here, the companies' key assets are the knowledge and the capabilities of its work force. The value to the customers lies in getting the right answer to their questions, not in the effort it takes the company to provide it.

[59] Zenger, T. R. and W. S. Hesterly (1997), "The Disaggregation of Corporations: Selective Intervention, High-Powered Incentives, and Molecular Units," *Organizational Science,* vol. 8, no. 3, pp. 209-222.
[60] Ghazal, Ahmad, Tilmann Rabl, Minqing Hu, Francois Raab, Meikel Poess, Alain Crolotte and Hans-Arno Jacobsen, *Bigbench: Towards an industry standard benchmark for big data analytics*, ACM, 2013.
[61] Stabell, Charles B, and Øystein D Fjeldstad, "Configuring value for competitive advantage: on chains, shops, and networks", *Strategic management journal,* 19, 5, 413-37.
[62] DNV GL AS, "DNV GL STRATEGIC RESEARCH & INNOVATION
POSITION PAPER 4-2014", 27.07.2014.

Examples would be the public and the health sector (see Section 1.4.3.2). The value network model provides a whole network to their customers who can conduct their own businesses using the network. Here, the main value lies. in the capabilities and size of the network and every additional customer joining enhances the overall value of the network. An example would be life sciences (see Section 1.4.3.2).

Business Model	Impact from big data	Examples (see Section 1.4.3.2)
Value chain	• big data enables further optimisation of operating value chains • can be used to move towards value network	retailing, manufacturing
Value shop	• Change of mode of operation • Problems that customers can be helped with • Efficiency and scalability of services, e.g. by efficiently re-using data and analytics	lifesciences
Value network	• Increase the overall value of the network • Capitalize on data generated by the network	public sector, health sector

Table 4: Business Models, i.e. value chain, value shop and value network, with examples (adapted from DNV GL AS[63])

Business Models in Various Sectors: Examples

In the following we investigate a number sectors where big data has been playing a role and might do so in the future. Our examples include retailing and manufacturing as private sectors, health care combining public and private sector, the public sector as well as life-sciences combining private and academic sector.

Retailing

In the future retailing might track individual behaviour of the customers with the help of online streams, model their behaviour and position itself optimally accordingly. Consider online booksellers being able to tie purchases to individual customers in a completely different way compared to brick and mortar booksellers[64]. Not only did online booksellers know what their customers buy they also know what they search for and can adapt their promotion strategies accordingly. Algorithms help them identify books and other products that the customer might wish to buy next.

[63] Ibid., p.33
[64] McAfee, Andrew, Erik Brynjolfsson, Thomas H Davenport, DJ Patil and Dominic Barton, "Big Data", *The management revolution. Harvard Bus Rev,* 90, 10, 61-67.

The use of big data in this context goes hand and hand with privacy issues[65]. Profiling of customers might lead to disturbing situations as the example of the pregnant teenage girl shows (see Section 1.1). Current examples of retailing using big data cover even brick and mortar companies, such as the U.K. retailer TESCO. TESCO gathers transaction data of its ten million customers via a loyalty card programme. Then, it analyses the data to take informed decisions on pricing, promotions, and shelf allocation[66]. The example of TESCO shows the problems with collecting and using data of individuals without them agreeing to this or even knowing what is happening to their data[67]. It quietly built a profile of people including a map of personality, habits in travelling and shopping etc. At the same time TESCO's subsidiary, Crucible, refused to disclose the data it holds while at the same time selling to other big groups, such as Sky, Orange and Gillette. The use of such data is manifold[68]: Some retailers, such as Fresh Direct, updates the prices and promotions daily or even more frequently based on data feeds from online transactions, visits by consumers to its web site and customer service interactions. Some other companies such as Ford Motor, PepsiCO and Southwest Airlines, determine the effectiveness of their promotion campaigns based on the analysis of their customers postings on social media sites such as Facebook and Twitter.

Companies originating on the Internet, such as Amazon, eBay, Google, have been the early leaders in using big data analysis to shed light on their customers' behaviour[69]. They have gathered intrinsic understanding of where to put a button or whether to put a pop up screen to determine what would increase the sales and user engagement. Financial institutions, such as Capital One, which have been into the game fairly early, continue to refine its methods of segmenting their credit card customers to then tailoring their products according to individual risk profiles. Capital One itself conducts 65000 tests each year to experiment with combinations of markets segments and new products.

Manufacturing Sector

In manufacturing Western companies have faced severe competition from emerging countries due to the productivity growth in these countries[70]. A promising way to deal with this challenge is using big data analytics. In manufacturing firms can build on thorough knowledge and experience with innovation and technological change as well as IT to empower the concurrent engineering process.

Globally, there are seven big data levers identified across the whole manufacturing value chain (see Figure 4 for more details):

[65] Smolan, Rick, and Christoph Kucklick, "Der vermessene Mensch", *GEO,* 08, August 2013, 80-98.
[66] Bughin, Jacques, and Michael Chui, "The rise of the networked enterprise: Web 2.0 finds its payday", *McKinsey quarterly,* 4, 3-8.
[67] The Guardian, *Tesco stocks up on inside knowledge of shoppers' lives*, 2005, http://www.theguardian.com/business/2005/sep/20/freedomofinformation.supermarkets.
[68] Bughin, Jacques, and Michael Chui, "The rise of the networked enterprise: Web 2.0 finds its payday", *McKinsey quarterly,* 4, 3-8.
[69] Ibid.
[70] McKinsey & Company, "Big data: The next frontier for innovation, competition and productivity", 2011, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

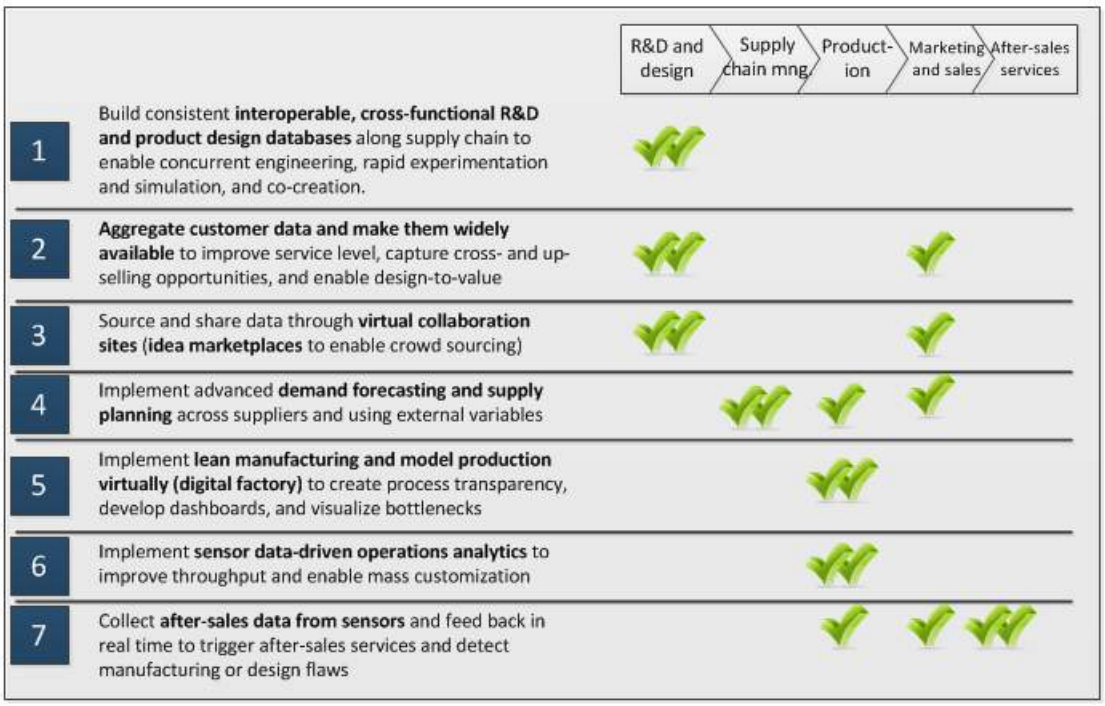


Figure 4: Seven big data levers across the whole manufacturing value chain[71]

Generally spoken big data increases the opportunities for using algorithms and machine mediated analysis[72]. For example, at some manufacturers production lines are self-regulated by algorithms analysing sensor data from production lines. The result is less waste and avoiding costly and potentially dangerous human interventions, thereby increasing productivity.

Products ranging from copiers to jet engines are now providing data streams that track their usage[73]. This means that manufacturers can analyse the incoming data and react to problems instantly – sometimes even in automated ways, e.g. by scheduling pre-emptive repairs before failures can disrupt customers' operations. All that leads to better performance and risk management, often based on insight remaining hidden without big data analytics. It is likely that more and more companies will implement these solutions as the prices of sensors, communications devices, analytic software and the likes will keep falling in the future.

Healthcare

In healthcare huge quantities of data are collected inside hospitals, primary care units, pharmacies as well as insurance companies and health authorities[74]. For an aging society facing more and more cases of diabetes, heart and lung diseases there are huge potential benefits of applying data analytics in healthcare. Data in healthcare comes as big data[75]: One

[71] Ibid., p.78
[72] Brown, Brad, Michael Chui and James Manyika, "Are you ready for the era of 'big data'", *McKinsey Quarterly,* 4, 24-35.
[73] Ibid.
[74] DNV GL AS, "DNV GL STRATEGIC RESEARCH & INNOVATION POSITION PAPER 4-2014", 27.07.2014.
[75] Ibid.

reason is its sheer volume. The other one is the complexity, diversity and timeliness of healthcare services. Analytics of big data are particularly relevant in the following areas of healthcare: diagnostics, patient monitoring and management in order to provide the right treatment at the right time, management of patient care processes from the first encounter with the general practitioner through all specialists, lab investigations, hospitals as well as medicament treatments.

Big data analytics might help dealing with the challenges of providing health care in our ageing societies. Examples are

- using unified electronic health records (EHRs). Compared to paper records the use of EHRs reduces management costs, medical errors, thereby leading to improved care, diagnosis and treatments.
- collecting huge amounts of information about patients, by using EHRs[76] imaging technology (CAT scans, MRI) in modern medicine or genetic analysis, such as DNA microarrays[77]. Already today medical researchers apply data mining to large data sets, thereby better understanding genetic and environmental causes of diseases and providing more effective means of diagnosis.
- using embedded and medical devices[78]. More and more sensors of all types, including such that can be implanted into the body will capture vital and non-vital biometrics, track the effectiveness of medication and correlate bodily activity with health. This will heavily increase the ability of the elder to live independently in their homes.

Public Sector

Big data might help governments around the world to increase their productivity[79]. In the aftermath of the recent global recession they have to spend public money to stimulate growth as well as take care of the increasing demands for medical and social services by an aging population. There are numerous big data levers. As data gets more ubiquitous, making the data available to stakeholders will create value. The canonical example of this sort of efficiency is the transparency in the public sector. In addition to sharing relevant data across governmental departments, using advanced analytics, segmenting customers into segments and tailor-making public services based on new algorithms would create tremendous value. Governments could reduce their costs of administration considerably if they fully exploit public sector data[80]. For the twenty-three largest European governments the OECD[81] estimate potential savings of 15% to 20%, equivalent to between EUR 150 billion and EUR 300 billion[82]. This estimation does not include additional benefits arising from greater access to and more effective use of public-sector information (PSI), such as obtained from improved weather forecasts, better traffic management, better crime statistics or improved transparency of government functions. The additional benefits of PSI is around EUR 32 billion in 2010.

---

[76] OECD, *Exploring Data-Driven Innovation as a New Source of Growth*, OECD Publishing, 2013.
[77] Bryant, Randal, Randy H Katz and Edward D Lazowska, *Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society*, December, 2008.
[78] Gantz, John, and David Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east", *IDC iView: IDC Analyze the Future*.
[79] McKinsey & Company, "Big data: The next frontier for innovation, competition and productivity", 2011, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.
[80] OECD, *Exploring Data-Driven Innovation as a New Source of Growth*, OECD Publishing, 2013.
[81] Ibid.
[82] McKinsey & Company, "Big data: The next frontier for innovation, competition and productivity", 2011, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

Big data might improve tax earnings of governments as well as public at the same time. Examples are

- making decision making of tax agencies easier[83]. Tax agencies have already been using data analytics to identify potential tax fraud in order to further investigate it. This means that the filtering takes place by using huge data bases and the actual investigation still takes place by human beings.

- smart grid technologies reducing or better managing electricity consumption[84]. The large volumes of data on energy and resource consumption patterns can provide information about consumption data and prices in real time. Moreover, this information can be used by consumers to adjust their energy and resource consumption to current production capacities. OECD[85] estimates that overall the use of smart grid technologies could reduce CO2 emissions by more than 2 gigatonnes (billion tonnes) by 2020, equivalent to EUR 79 billion. While smart grid technologies offer large opportunities they also require substantial investment in supporting infrastructure in order to connect billions of devices[86]. Connecting the next billion smart devices to the Internet and the corresponding exchange of exabytes of data every month will challenge the operation of current communication infrastructures. Moreover, smart grid technologies come with substantial regulatory challenges, particularly regarding security and data privacy[87], surveillance footage used in crime investigations[88]. As usually date, time, location, etc. is automatically attached to video files in surveillance footage they are already today very helpful in crime investigation. The more intelligence is embedded the cameras the easier it will be to capture, analyse, and tag the footage in real time, thereby speeding crime investigations.

Life Sciences

Big data might also change the way of innovating. A prominent example are life sciences, particularly bioeconomy[89]. DELSA Global (Data-Enabled Life Sciences Alliance International) (http://www.delsaglobal.org/) brings together stakeholders of life sciences. Starting point of their concept is that there is an enormous chunk of the data collected that is underutilised. In order to make better use of this data DELSA Global wants to move from the "single scientist- single project" model to collective innovation processes by utilizing online platforms and distributed computing relying on big data in life sciences. They do so by building and maintaining a network of professional societies, funding agencies, foundations, companies, and citizens on the one hand side and life science researchers and innovating agents in computing, infrastructure and analysis. By bringing them together DELSA Global aims at translating new discoveries into tools, resources and products.

---

[83] Stoyanov, Veselin, and Claire Cardie, *Topic identification for fine-grained opinion analysis*, Association for Computational Linguistics, 2008.
[84] OECD, *Exploring Data-Driven Innovation as a New Source of Growth*, OECD Publishing, 2013.
[85] Ibid.
[86] Ibid.
[87] DNV GL AS, "DNV GL STRATEGIC RESEARCH & INNOVATION POSITION PAPER 4-2014", 27.07.2014.
[88] Gantz, John, and David Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east", *IDC iView: IDC Analyze the Future*.
[89] Kolker, Eugene, Elizabeth Stewart and Vural Özdemir, "DELSA Global for "Big Data" and the bioeconomy: catalyzing collective innovation", *Industrial Biotechnology,* 8, 4, 176-78.

2.4.4 Data Warehousing, Service Models, and Data Markets

In this section we briefly review the economic and entrepreneurial opportunities afforded by computing technology. These opportunities are best understood in terms of a widening sphere of potential impacts. The first order opportunities involve the sale of computer hardware. The second order opportunities involve computer services, and the third order opportunities involve the sale or management of opportunities. The section is organized to discuss this by first discussing the first and second order effects – the data warehouses and cloud computing services. Then we discuss the third order effects enabled by new computing services – the collection and re-sale of data.

Existing and newly emerging companies can create new business and revenue streams[90]. For that they use the huge volumes of data and of capacities to store and process them. The companies provide services and products ranging from providing raw data to fully implementing tailor-made solutions for their customers. As a consequence new data markets emerge differing from traditional markets. The potential value of these big data induced markets is large. To give an idea of the value that big data can create Mc Kinsey and Company estimated the potential annual value of big data for the European public sector administration €250 billion and $300 billion for US health care sectors.

An even bigger economic impact may result from new computing service models, or "the cloud." Cloud computing, and related "software as a service" business models are estimated at $72 billion world wide revenues in 2015. These services are variously expected to grow between 20% and 30% yearly at least until the year 2018. This compares with 5% yearly growth in regular information technology acquisitions.[91]

2.4.5 Data Warehousing and Other Computing Service Models

At the core of modern big data enterprises is a capacity to store data in databases, for instance in a data warehouse. Also core to the activity is to collect and distribute information (for instance with web servers). Other important tools in the ecology are hardware for transmitting the data, for processing the information (with high performance computing clusters) or for filtering and analysing the information to gain insight (with middleware or other software). The big data computing ecology is discussed more thoroughly in the first work package of the BYTE project.

Data warehouses and analysing the data collected can help determining the effectiveness of their pricing strategies and promotion campaigns and better manage their inventory and supply chains. Hewlett Packard constructs a data warehouse for Wal-Mart to store 4 petabytes (4000 trillion bytes) of data, representing every single purchase recorded by their point-of-sale terminals (around 267 million transactions per day) at their 6000 stores worldwide[92].

---

[90] DNV GL AS, "DNV GL STRATEGIC RESEARCH & INNOVATION POSITION PAPER 4-2014", 27.07.2014.

[91] Colombus, L. (2015), "Roundi[of Cloud Computing Foreasts and Market Estimates, 2015," Forbes, Accessed online 17 July 2015: http://www.forbes.com/sites/louiscolumbus/2015/01/24/roundup-of-cloud-computing-forecasts-and-market-estimates-2015/

[92] Bryant, Randal, Randy H Katz and Edward D Lazowska, *Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society*, December, 2008.

An important player in data markets on top of producers/analysts of big data and their consumers are employees. In particular in data management and analytics there are considerable mismatches between the supply of and demand for skills[93]. This might not only slow down the adoption of big data analytics but might also lead to missed chances of creating new jobs. It would require a multidisciplinary approach to education, training and skills development in order to overcome this problem.

2.4.6 Data Markets

The term "data market" resembles a traditional market place where economically significant, sought after data can be traded between data suppliers, who harvest data on a particular matter, and data consumers, who require data on this particular matter. Typically one envisions this trading to be performed in a business to business form. Indeed, data markets are trending among large and medium scale companies nowadays. Companies such as Azure Data Marketplace, salesforce.com, InfoChimps.com, datamarket.com sit on one end of this trading table, whereas various companies that crave for data to distil insights out of data sit on the other end.

Vom Lehm[94] offers a comprehensive evaluation of data markets and privacy. Of particular relevance here are the various alternative data market proposals which have been suggested. These include the national information market of Laudon[95], a market proposal by Schwartz[96] which preserves hybrid inalienability, the personal information market of Novotony and Spiekermann[97], and the humanistic information economy as presented by Lanier[98]. Laudon[99] argues that there ought to be a national information exchange, driven by personal deposits of information rights, and paid for with corresponding rents on the data by organizations which use the personal data. Schwartz[100] critiques the centralized character of the national information market, and endorses a highly distributed institution driven by specific patterns of use, and corresponding liabilities for suspected violations of agreed data use policies. The personal information market of Novtony and Spiekermann[101] is explicitly tailored for a European institutional context. This more detailed market proposal offers multiple tiers of service, with tailored information hiding and withholding by user and application need. As a final note, the market proposition of Lanier[102] explicitly preserves the rent for the creative classes in society; it stands in staunch opposition to markets created on behalf of large organizations and their silos of information.

Vom Lehm[103] distinguishes between primary and secondary data market. The primary data markets directly interface with users, while the secondary markets aggregate the data without

[93] OECD, *Exploring Data-Driven Innovation as a New Source of Growth*, OECD Publishing, 2013.
[94] vom Lehm, H, *On data markets as a means to privacy protection: An ethical evaluation of the treatment of personal data as a commodity*, Delft University of Technology.
[95] Laudon, K. C. , "Markets and privacy", *Communications of the ACM,* 39, 9, 92-104.
[96] Schwartz, P. M., "Property, privacy and personal data", *Harvard Law Review,* 117, 7, 2055-125.
[97] Novotny, A., and S. Spiekermann, *Personal information markets and privacy: A new model to solve the controversy*, Gabler Verlag, 2013.
[98] Lanier, J. , *You Are Not a Gadget: A Manifesto*, Alfred A. Knopf, New York, 2010.
[99] Laudon, K. C. , "Markets and privacy", *Communications of the ACM,* 39, 9, 92-104.
[100] Schwartz, P. M., "Property, privacy and personal data", *Harvard Law Review,* 117, 7, 2055-125.
[101] Novotny, A., and S. Spiekermann, *Personal information markets and privacy: A new model to solve the controversy*, Gabler Verlag, 2013.
[102] Lanier, J., *Who Owns the Future?*, Simon & Schuster, San Jose, 2013.
[103] vom Lehm, H, *On data markets as a means to privacy protection: An ethical evaluation of the treatment of personal data as a commodity*, Delft University of Technology.

directly remunerating the original user. In practice, the current big data environment involves a set of bilateral contracting arrangements with various service providers. Hence it less resembles a market than a contract. And hence, it affords whatever terms and conditions tolerated by the public and permitted by national and international law. Vom Lehm describes this regime as the free service status quo[104]. The free service status quo may be evolving to a highly distributed model of open contracting where more economically valuable data is collected, yet the data is marshalled for highly specific application in a limited use setting. See for instance Ethereum, a second generation bitcoin platform, as an example of this[105].

Data trading is powered by the Internet, which brings in all the advantages of it, such as crowd sourcing, massive exploitation and exploration abilities and so on. The ample possibilities of the Internet make it possible to create unprecedented ways of economic relations in contrast to the traditional sense of data markets. Many companies prefer to hire a third party that assembles and curates datasets or creates insights out of proprietary data[106]. Outsourcing data management is an example of developing economic change around data. The business model of Kaggle is an example in this regard[107]. Kaggle provides the data problems of the companies to a large base of data scientists (which can be any one with data science capabilities or someone who would like to acquire these capabilities.) who analyse the data to solve data problems for the companies to win some awards. The currency in data markets is not only money but can be data at times. Salesforce.com, which is a cloud based company that is best known for its customer relationship management (CRM) product, created a data market where individuals and organizations give away their contact information in exchange for contact information data of the other organizations and individuals.

2.5 CONCLUSIONS

Big data will affect all walks of life in industries, science and private relationships and comes with considerable privacy issues because of new data sources and actors as well as the increasing ease of linking and processing data[108]. The challenge is to realise economic and social benefits while at the same time effectively protecting the privacy of individuals.. However, due to different value potential and abilities the economic effects will differ substantially for the various stakeholders. Big data is promise and challenge at the same time. While it promises huge productivity and welfare increases it also comes with numerous issues. For some of them solutions might be available – sometimes only temporarily because of ever changing technologies and application areas. To give an example, privacy issues are of great concern to private parties[109] and economic and technological solutions will have to find ways of incorporating privacy solutions[110].

---

[104] Ibid.
[105] Schneider, N., *Code your own utopia: Meet Ethereum, bitcoin's most ambitious successor*, 2014.
[106] Elbaz, Gil, *Data Markets: The Emerging Data Economy*, 2012, http://techcrunch.com/2012/09/30/data-markets-the-emerging-data-economy/.
[107] Kaggle.com
[108] Ibid.
[109] Smolan, Rick, and Christoph Kucklick, "Der vermessene Mensch", *GEO,* 08, August 2013, 80-98.
[110] Sogeti, "No More Secrets with Big Data Analytics", Groningen, 2013, http://www.google.de/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=1&ved=0CD0QFjAA&url=http%3A%2F%2Fvint.sogeti.com%2Fwp-content%2Fuploads%2F2013%2F11%2FSogeti_NoMoreSecrets.pdf&ei=6zqwU4WvJImw7AbysoHQCg&usg=AFQjCNFMAmRVE6v4loN3Kjsg7Vp_sBguKg&bvm=bv.69837884,d.bGE.

Big data is expected to be a new type of asset to both the private and the public sector as it, represents a key basis for competition[111]. All stakeholders are challenged to equip themselves with big data and analytics in order to be able exploit the potential of big data. This does not only hold for companies but also for governments. In fact, big data might make the difference in how countries and companies, compete and thrive. Analytics of big data might offer solutions to problems of a struggling global economy. However whether they will be able to do that is questionable[112]. The companies do not only have to acquire new capabilities but also to develop new perspectives on how to operate in the big data era.
All stakeholders have to deal with big data responsibly and carefully. "Self-interests of individuals, companies or institutions have limits, where the public interest is affected, and public interest is not a sufficient justification to violate human rights of individuals."[113]. Many of the problems we face in the big data era result from problems of the Internet, more particularly from Web 2.0 and other applications [114]. As the internet was not originally designed for Web 2.0 the current technical solutions do not suffice. A way out might be organizing the Internet by using concepts of social self-organization[115].

Whether or not big data will bring more chances or problems is not only an economic but a much broader matter. Many of the economic issues of firms and of other stakeholders discussed here, such as copyrights, privacy, compliance with regulations, relate to legal, political, social and ethical as well as the political issues. These will be discussed in Section 3 to 5, i.e. deliverables in the work packages.

[111] OECD, *Exploring Data-Driven Innovation as a New Source of Growth*, OECD Publishing, 2013.
[112] DNV GL AS, "DNV GL STRATEGIC RESEARCH & INNOVATION
POSITION PAPER 4-2014", 27.07.2014.
[113] Helbing, Dirk, and Stefano Balietti, "From social data mining to forecasting socio-economic crises", *The European Physical Journal-Special Topics,* 195, 1, 3-68.[p. 23]
[114] Ibid.
[115] Ibid.

3 LEGAL ISSUES IN BIG DATA

Hans Lammerant, Antonella Galetta and Paul De Hert,
Vrije Universiteit Brussel

3.1 OVERVIEW

Big data is not a clearly defined phenomenon, but it is clear that it has a strong impact on science, on how business is done, on the technical infrastructure sustaining daily life, on social relations, etc. With the emergence of search engines like Google or social media like Facebook and Twitter big data entered daily life. Big data is not a technique only used within a clearly defined domain, but it interferes now with and even mediates a wide range of social relations. By consequence, big data also touches upon the legal frameworks regulating these social relations.

It is not so obvious to phrase this in terms of externalities. We could define the objectives of these legal frameworks or the values protected by them as externalities and evaluate the effects of big data processing on these values and on the effectiveness of the legal frameworks when confronted with big data processing. But this would lead to a too negative perspective when technological change put legal frameworks under pressure, as in that case it is clear that these frameworks are 'negatively' impacted. Such perspective would forget that big data processing brings along strong advantages in terms of economic efficiency and new capabilities. Positive and negative effects are often intertwined. The better question concerning legal frameworks is if that are still well-adapted to the technological environment in order to reach their objectives.

These legal frameworks, like intellectual property rights (IPR) or data protection, were shaped in another technical environment. The exemptions present in the intellectual property regime are mainly directed to a print environment. The IPR framework has been updated, but it is clear that such updating still struggles with the new digital environment. And the data protection framework of directive 95/46 originates from the early computer age where databases were still small, clearly separated and could be easily located. Also this framework gets updated with the new proposal of General Data Protection Regulation (GDPR). But again, the question remains if this update can adequately deal with big data. In both cases the question is if updating is enough or that a new conceptual framework for regulation is needed.

Two important legal frameworks in the context of data processing are intellectual property rights (and the contractual framework to allow their use and transfer) and data protection. IPR have a clear economic objective. It grants temporary monopoly rights in order to allow an adequate return for investors. Such monopoly rights are an impediment to the free exploitation of the protected goods by others and lead to higher prices. But the objective is to strike the right balance, which would lead on the macro-level to higher investments and more economic growth. Question is if this balance is indeed still present or that the IPR framework erects too much barriers obstructing the development of the data economy. Data protection has a totally different objective and aims to protect privacy as a fundamental right and an important social value. Therefore it regulates the use of the core element used by big data processing, data, when it is related to identifiable persons. In this case a balance has to be struck between the protection of privacy and giving enough space for all sorts of societal and economic processes involving personal data.

Data gets processed in order to inform all sorts of decision making. Therefore big data processing can have a much wider impact than on the legal frameworks directly dealing with data. An important one is non-discrimination legislation, which forbids to ground decisions on certain elements like sex, race, sexual preference, etc. But the impact is even wider. Legal frameworks consist of general terms, which need to get a specific meaning in their actual application. Technological evolution can influence the scope of such legal frameworks. For example, the definition of personal data depends 'on the means reasonably likely to be used' for identification. In other words on the technological capacity to identify persons from data, which clearly changes in the context of an abundance of data and the development of data analysis techniques capable of dealing with big data. Such changing meanings of general terms and their consequences can be very widespread. For example, investigation measures in criminal or administrative law are linked with notions like reasonable suspicion. Such suspicion gets constructed in new ways in the context of big data. Tax administrations link different data sources and use data mining to uncover taxpayers or companies which are outliers compared to the normal patterns. Being an outlier in the statistical sense can in certain circumstances be a building block of suspicions which trigger an investigation. Or certain correlations with suspicious activity can be enough to widen this suspicion. 'Wrong time, wrong place' is an experience which can happen in a lot more ways. Similar evolutions can be seen in private law contexts with notions like risk. Especially with new technological developments such interpretation of general terms can be tricky and legal uncertainty abounds if not complemented with extra mechanisms to provide further guidance. Regularly updating laws or soft law approaches like advices by DPAs or professional organisations can help to shorten the learning curve in dealing with new circumstances and restoring a predictable legal environment. But for big data we are still in the middle of this learning curve. This changing nature of the construction of suspicion points to a widespread changing nature of decision making. It raises the problematic of due process, or the capacity to influence, have a voice in or to object to decisions made about oneself. Decisions about a person are based on certain information or data, but in the big data context the grounds for a decision can become very difficult to trace. The large amount of data from diverse sources combined with the opaque character of most data mining techniques makes such tracing often impossible. Further, the spreading of personal data also makes it difficult for a data subject to know where data about him is being used and by consequence to know where decisions affecting him are made. All this can make it difficult or impossible for an individual to have a voice in decisions made about him and to object and correct certain assumptions they are based on. This while decisions become based on a wide range of fuzzy data and on correlations which lead to certain assumptions, without being actually true. Individuals can get locked up in certain profiles, without having the instruments to change or influence the assumptions on which they are based. This can even lead to unintended discrimination. Due process is therefore an important issue to sustain and restore the autonomy and freedom of persons. The capacity to make one's own decisions is a core element of responsibility.

A last important issue is the impact of the virtualization processes underlying big data processing on the functioning of basic legal mechanisms. Legal frameworks are often implicitly based on natural or material characteristics linked to human practices. Digitalization has led to abstraction processes from material carriers in a lot of areas. This led to much easier communication in terms of distances and amounts of information, and has led to a rearranging and reorganizing of a lot of practices and business processes. This also has put basic legal mechanisms under strain, like the geographical division through jurisdiction which defines applicable law and competent courts, and liability mechanisms which

determines responsibility among actors. Big data processing is not the first or sole cause of this, but a further step augmenting the challenge to these legal mechanisms.

The legal frameworks mentioned so far are not specific directed to big data. They concern all data, big and small. The term 'big data' originally points to a technical difference in dealing with data: the 3 V's (volume, variety, velocity) necessitating other techniques than the traditional SQL-database on a single computer. But from a legal point of view it does not make a difference if data is stored in a MySQL-database on a single computer or managed with a NoSQL-database over Hadoop on several computers in a cloud. The legal frameworks mentioned get triggered by other qualifications of the data: does it concern personal data, is it the expression of an author, etc. But these legal frameworks applicable on data in general can pose specific problems when applied to big data processing. To consider the specific application of these legal frameworks to big data we looked into certain aspects of big data processing which makes the implementation of these frameworks challenging: the use of data from many sources, data analytics and more specifically data mining as form of processing, and for the liability and jurisdictional issues also cloud computing. The use of data from a wide range of sources can also the rights of a wide range of persons and create practical difficulties in implementing the protection of their rights (like transparency and access rights for data subjects when personal data is used, getting authorizations from copyright holders, etc.). This exacerbates when the data originates from many jurisdictions and triggers different laws into application. New forms of data analytics like data mining deal with a large amount of data and similarly can trigger these right protection mechanisms. If these protection mechanisms are well-adapted to this use of data or need to come into application at all is questionable. Further the virtualization processes underlying big data processing and cloud computing as enabling technology can lead to opaque and complicated architectures. Also the technological convergence of services leads to problems in the application of legal frameworks, as these often were conceptualized on distinct use cases which now get blurred. This creates difficulties for the application of liability mechanisms, while it also creates jurisdictional problems, including a risk for the inflation of applicable laws. As a conclusion we can state that the existing legal frameworks applicable on the use of data in general pose specific problems for big data and are not well-adapted to novel technologies. In the following sections we discuss each of these issues more in depth.

3.2 INTELLECTUAL PROPERTY RIGHTS, LICENSING AND CONTRACTING

3.2.1 Intellectual property rights

Copyright protection to data sources results in complications for big data processing. Before the processing it is needed to get authorization of the right holders on protected sources, which can be a major hurdle in use cases combining a wide range of sources and lead to large transaction costs. IPR protection of data in its existing shape results in data enclosures and barriers stifling big data processing.

Major starting questions are when do intellectual property rights exist over datasets, and which uses are in such a case exclusive to the right holder and which not. Copyright can exists over the individual data as well as over the database as a whole, while also a sui generis-protection exists for databases. If a copyright exists for a certain dataset, is not very easy to determine. Copyright is linked to an originality-requirement. And the protection differentiates between the individual data items and the database, which each have to fulfil this requirement. Copyright of individual data items grants exclusive rights on the individual item. Copyright

protection for databases results from the copyright for collections. The Berne Convention already protected collections of literary and artistic works which constitute intellectual creation by reason of selection and arrangement of their contents, such as encyclopaedias and anthologies.[116] The WIPO Copyright Treaty and TRIPS Treaty provide an explicit copyright protection for databases. It is linked to the organisation and structuring of the data but does not extend to the individual data items itself. General principle of copyright is that it protects expressions, but not ideas in itself, nor procedures, methods or mathematical concepts.[117] The Berne Convention also states that the protection does not apply to "news of the day or to miscellaneous facts having the character of mere items of press information".[118] Aim is to protect products of human intellect and creativity. Trigger for the protection is therefore some sort of originality. The threshold for this protection gets defined slightly different across countries. Common law countries have interpreted it as originating from the author[119], while civil law countries generally required a slightly higher element of creativity as part of the originality requirement[120]. But the differences have been small and in general the threshold is low, while also the US courts required a minimal amount of creativity. In the Feist decision the US Supreme Court rejected copyright protection for an alphabetic telephone directory. For datasets this has an important consequence: not all data is protected by copyright, but only those which meets the originality-requirement. Purely factual information is not protected under copyright. For instance, also maps have been subject to copyright controversies, as the factual geographical information as such lacks the element of creativity. Maps are expressions using geographical information and also collections of such data, but when the map is a too obvious reproduction lacking creative choices it is not protected by copyright.[121] Also, even when copyright protection is given to the expression (like a map or a table with data) it does not extend to the underlying facts. The factual information can be used by others, as long as they do not reproduce it in the protected expression.

The EU also introduced a second sui generis protection of databases, which protects databases based on the investment made in them. Such a protection does not exist in the US and is not contained in the intellectual property treaties. Database protection is provided by directive 96/9/EC of 11 March 1996 on the legal protection of databases. This directive contains 2 forms of protection for databases, one as copyright, another as a sui generis right. These protections can coincide. The sui generis-right protects the maker of a database "which shows that there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents".[122] What is protected is the economic investment in creating the database. To assess what is a substantial investment, the cost for obtaining, creating or updating the individual data items cannot be taken into account. Only the costs associated with the actual making and maintenance of the database itself are relevant.[123] The maker of the database is given the right to prevent extraction and re-

[116] Berne Convention for the Protection of Literary and Artistic Works (Paris Text, 24 July 1971), art. 2 §5.
[117] WIPO Copyright Treaty (Geneva, 20 December 1996), art. 2; Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) (Marrakesh, 15 April 1994), art. 9 §2.
[118] Berne Convention, art 2 §8.
[119] Paul Goldstein, *International Copyright. Principles, Law, and Practice*, Oxford University Press, New York, 2001, p. 161.
[120] Ibid., p.164.
[121] Janssen, Katleen, and Jos Dumortier, "The Protection of Maps and Spatial Databases in Europe and the United States by Copyright and the Sui Generis Right", *J. Marshall J. Computer & Info. L.*, Vol. 24 No.195, 2006, pp. 207-211.
[122] European Parliament and the Council, Directive 96/9/EC of 11 March 1996 on the legal protection of databases, art. 7 §1.
[123] Truyens, Maarten & Patrick Van Eecke, "Legal aspects of text mining", *Computer Law & Security Review*, Vol. 30, no. 2, 2014, 160.

utilization of the whole or of a substantial part of the contents of that database. Extraction stands for "the permanent or temporary transfer of all or a substantial part of the contents of a database". Re-utilization is any form of making available to the public, like the distribution of copies or renting by on-line or other forms of transmission. This right does not prevent lawful use, consisting of extracting or re-utilizing insubstantial parts of database contents. The substantiality can be assessed both quantitatively or qualitatively. Further may this use not conflict with the normal exploitation of the database or unreasonably prejudice the legitimate interests of the maker of the database. This lawful use does not affect other limitations posed by copyright or other rights residing in the contents. The sui generis-right provides less exceptions than copyright, but possibly relevant is the exception for non-commercial scientific research.

The EU introduced this sui generis protection based on the assumption that property rights, which come down to temporary monopoly rights, attract investments and therefore stimulate the economy. In the US such protection was refused on a similar economical reasoning. Legal protection based on mere effort or investment has been refused by the Supreme Court in the Feist decision. The Supreme Court took distance from court decisions which also granted protection to 'sweat of the brow' or 'industrious collection', through which courts had earlier developed a protection for factual collections. Instead it made clear that originality was an essential requirement and that facts or factual compilations could therefore not receive copyright protection. The Court grounded that requirement on the objectives of copyright protection listed in the Constitution "to promote the Progress of Science and useful Arts". Copyright also needs to allow others to build upon the ideas and information contained in a work, which is the rationale for only granting protection to the expression but not to facts. Ian Hargreaves pointed in his review of the intellectual property framework, which centred on the question if the existing framework was still promoting innovation and growth, to the evaluation in 2006 by the European Commission of the Database directive. This evaluation shows less investment instead of growth, while the US market kept growing without such protection.[124] Hargreaves sees this as an example of policy development inconsistent with the available evidence.[125] The European Commission has kept the directive unchanged seen the large support of the concerned industry for the directive. It can be questioned if such large support shows the economic value of the directive in general or that it shows the value for a specific interest group. More in general the question can be raised if a data economy would not be better off with less protection of databases through IPR. In fact, the availability and usability of data leads to positive network effects. Positive network effects imply that an extra user augments the value of the good for all other users. This is a known feature of communication equipment (e.g. the first fax was useless, but got more usable once more fax machines got into use) and other things used in networks. Although less obvious, a similar phenomenon can be observed in the linking of data sources. A data source of good quality which is open will be used in combination with other sources and for other analyses then the ones originally envisaged. When new sources become available this enlarges the options for re-use of the older sources. This is augmented by the fact that the marginal cost of such new re-use is near zero, at least when this marginal cost is not artificially raised through IPR.[126] Stimulating economic growth is in this case better served with opening up data sources, even

[124] European Commission, *First evaluation of Directive 96/9/EC on the legal protection of databases*, DG Internal Market and Services Working Paper, 12 December 2005, pp. 22-23.
[125] Hargreaves, Ian, *Digital Opportunity: Review of Intellectual Property and Growth*, May 2011, http://www.ipo.gov.uk/ipreview-finalreport.pdf, p.19.
[126] A similar mechanism was explained for software by Benkler, Yochai, *The wealth of networks: How social production transforms markets and freedom*, Yale University Press, 2006.

if it can hinder traditional business models and pushes for new ones (like happened around open source software).

Copyright gives the right holder the exclusive right to authorize specific uses of the protected work. It generally reserves the right of reproduction, adaptation and distribution to the right holder, as well as the right to be credited as the author. Access is as such not one of the rights reserved to the copyright holder, but the Berne Convention reserves the right to authorize broadcasting or the communication to the public of their works over a range of channels or means of communication.[127] Most copyright legislation limits these rights by formulating exceptions, like the right to quote or to parody, or fair use. The fundamental problem with the copyright framework is that its basic concepts are based on the physical and technical limitations from the print age. One of the problems is what constitutes 'reproduction', which is one of the actions on which the author has exclusive rights. Concerning printed matter a clear distinction can be made between use of the work by reading or looking and reproduction of the work by copying or printing another physical exemplar. This distinction cannot be made so easily in the digital age, where using often involves the making of another copy of the data. For instance, reading a webpage consists of sending data to other computers which reconstruct a copy of the file similar to the one present at the sending computer. In a following step this local copy is used to render the webpage in a browser and making it human readable and visible. What is use and what is reproduction does not follow automatically any more from the technical features but this distinction has to follow from the legal distinctions made between copies and the underlying technical reproduction processes. Copyright law has been updated to allow technical reproduction during communication and similar technical processes. But till now it has been done in a restrictive manner. Also courts make divergent applications of this law, which often shows a lack of understanding of the technical processes. Text and data mining often starts with the collection of a large amount of text or other data, in order to process in the next stage. Technically such collection of data involves making a local copy. Is this in legal terms use of data as it is published on the internet, or is this a reproduction? Court decisions concerning web scraping are very divergent and show that technical processes are barely understood. Often the collection and making of a local copy for later analysis is confused with the reproduction in a new webpage and communicating of this page to the public. For instance, the Belgian Copiepresse-case states there is reproduction from the moment there is storage of data. But while it considered the reproduction and the communication of the data Google collected through Google cache and Google News a violation of the copyright of the authors, it had no problem with Google Search as such.[128] Basically because the court seemed to think that such search was an immediate referencing not involving any storage. This is in contradiction with the fact that Google cache is a local copy used for the data mining which builds the index. It is therefore necessary for the technical process. What is not necessary is giving the public an immediate access to that copy, which can be considered as a communication to the public. Several authors have pointed to the legal problems for text and data mining caused by IPR and policy discussions have been taking place in the EU and elsewhere.[129] They signal that the

[127] Berne Convention, art. 11 §1.
[128] Court of Appeal Brussels, 5 May 2011, *Google Inc. v. Copiepresse et al*, http://jure.juridat.just.fgov.be/oldf/view_decision?justel=F-20110505-18&idxc_id=252985&lang=fr
[129] Hargreaves, Ian et al, *Standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining. Report from the Expert Group*, Publications Office of the European Union, Luxembourg, 2014, http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf; Triaille, Jean-Paul, Jérôme de Meeûs d'Argenteuil and Amélie de Francquen, *Study on the legal framework of text and data mining (TDM)*, March 2014, http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf; Truyens, Maarten & Patrick

exceptions in the IPR framework are too narrow to allow data mining and have proposed to adapt the exceptions in the framework. More fundamentally, these legal problems for text and data mining resulting from copyright law can be seen as a result of considering data collection and the making of a local copy as a form of reproduction and not as an act of licit use.[130] It could be hoped for that courts come after a learning curve to a more stable interpretation in line with the technical developments. Legal changes like introducing a new exception for data mining, enlarging the exception for technical processes or making copyright law more technology-neutral through fair use-clauses can help shorten this learning curve. On the other hand, the fundamental problem of an outdated conceptualisation present in copyright law will remain and pose problems with other technological developments as well.

The difficulties posed by the copyright regime can be illustrated with user-generated content. That is content produced by users of internet services like Twitter and Facebook, Wikipedia, comments under newspaper articles or internet discussion fora, etc. The user as author has the copyright on his text, even if it consists of only one or a few sentences. Big data companies providing internet services, like Twitter and Facebook, clarify the legal situation by stipulating in their Terms of Use (with which the user agrees before being able to use the service) that the user grants a license to them. This includes the right to sublicense to other users. E.g. Twitter's Terms state that the user grants it "a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods (now known or later developed)".[131] Without such license through the Terms of use, an authorization has to be obtained from each of the individual users before their content can be 'reproduced' and in case no use of exceptions can be made. Getting such authorizations forms a large burden before the user-generated content can be used in big data processing. Examples of use cases are topical or sentiment analysis of comments on internet fora or newspaper articles. An added difficulty is when the user cannot be identified, in which case we have to do with an 'orphan work'. In general such orphan work becomes unavailable for protected uses.[132] In general we have the same problem as with data and text mining. The existing exceptions prove to be too limited to give a certain escape route for the burden of obtaining permissions. A specific exception for text and data mining or a more general fair use framework of exceptions would resolve a lot of problems.

The protection by copyright and sui generis database rights of data sources clearly limits big data processing. It sets up isolated data sources, and making them available involves high transaction costs due to obtaining the necessary licenses for each source. Restricting the protection would give space for data flows and combination of data sets, as well as for new

---

Van Eecke, "Legal aspects of text mining"; Ian Hargreaves, *Digital Opportunity: Review of Intellectual Property and Growth*; Australian Law Reform Commission, *Copyright and the Digital Economy. Final Report*, ALRC Report 122, 30 November 2013.

[130] Hargreaves, Ian et al, *Standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining. Report from the Expert Group*, pp. 53-54.

[131] Twitter, *Terms of Service*, version June 25, 2012, art. 5, https://twitter.com/tos. Facebook has a similar stipulation: Facebook, *Statement of Rights and Responsibilities*, version November 15, 2013, https://www.facebook.com/legal/terms, art. 2. Wikipedia stipulates that its users agree to submit content under the Creative Commons open content license CC BY-SA 3.0, see Wikipedia, *Terms of Use*, version June 16, 2014, http://wikimediafoundation.org/wiki/Terms_of_Use#Our_Terms_of_Use.

[132] The regulation of the use of orphan works by libraries and educational establishments provided by Directive 2012/28/EU of the European Parliament and of the Council of 25 October 2012 on certain permitted uses of orphan works does not apply on such user-generated content.

uses like data mining. This comes down to redrawing the balance on which the IPR framework is based in order to let it better achieve its objectives.

3.2.2 Licensing and contracting

Intellectual property rights do provide a default property regime regulating access to and control of data. This default regime only establishes a starting point from which can be deviated in contractual agreements. Private regulation through contract law by holders of intellectual property rights further shapes the possibilities and barriers to access data. Licensing is a mechanism through which private actors can shape the data landscape and eliminate or erect barriers for access.

Several practices are visible. An important one is departing from the IPR framework by turning products into services and making them available through restrictive licenses. This mechanism of end user license agreements (EULA) is best known for software, but can also be used for sources of data like online databases. It is often much more restrictive than the traditional intellectual property law. For instance, exceptions available in the IPR framework can be put aside by contract. Therefore, the proposals for new exceptions for data mining mentioned earlier often include to make this exceptions mandatory and protected against overriding through contract.[133] A totally different approach is the use of licenses to open up content. These open content licenses also use contractual mechanisms to deviate from the default rules of intellectual property law, but this time to guarantee access.

All these practices do suffer of problems linked to the law of contracts. The legal qualification of a license is not always clear, while this qualification has important consequences. Main question is if the license remains in the intellectual property regime or if it becomes a contract.
A license remains in the intellectual property regime when it does not contain new obligations for the user. In this case a license is a unilateral permission by the right holder/licensor to the licensee to use certain rights. The licensor promises not to claim specific rights against the licensee when the licensee reproduces or adapts the protected product in the circumstances specified by the license. Such specifications can limit the authorized uses or the group of licensees, as long as it does not depart from the default regime of copyright law. In this case a violation of the license is a copyright infringement and can be enforced against everyone based on IPR laws. When the license departs from the default law and includes new obligations (e.g. a choice of forum), it becomes a contract. This has important consequences: it needs to be accepted by the licensee and it is only binding on the parties involved in the contract. A violation of the license by the licensee will be a violation of a contractual obligation and be judged according to the law on contracts.

The legal status of licenses has to be checked in each jurisdiction according to the local law. And contract law is much less harmonized than IPR. By consequence this creates difficulties when licenses have to function in several jurisdictions or to determine the legal interoperability of data sources. Seen that a contract is an agreement between several parties, the validity of a contract depends on how local contract law recognizes such agreement, if certain formalities are required and on which consequences are given to this. Licenses on websites often consist of a long legal text and the possibility to click yes or no (so-called

[133] Hargreaves, Ian, *Digital Opportunity: Review of Intellectual Property and Growth*, p.51.

clickwrap licenses[134]). Question is if such clicking is considered as constituting an agreement to a contract, and if so to what extent. Often it is considered as an adhesion contract and the agreement is considered to be limited to the basic obligations linked to the transaction, but not to the whole license in all details. These differences in contract law make it unclear how the validity and scope of a license will be considered by the courts in different jurisdictions and how the license can be enforced. Case law from several jurisdictions show that courts often found that licenses could not qualify as a valid contract and were by consequence not enforceable, or only for a limited set of its terms and conditions.[135]

From this follows that a webpage with Terms of Use, without that any agreement is asked, is an even more weak attempt to private regulation, which is likely not enforceable. Large scale data collection practices, through screen scraping or web crawlers, are sometimes forbidden in such terms of use. When acceptance of such terms of use is needed and cannot be proved, data collection can only prevented if it is possible to invoke protection of intellectual property rights. So in general attempts at private enclosure of data proves to be difficult, or at least not straightforward.

Open content licenses try to avoid the difficulties connected with contract law by avoiding to include contractual obligations. They are often bare licenses remaining in the field of copyright law and are unilateral permissions given to use the licensed product. Conditions are built in as specifications to who is given such authorization. But if this is possible depends on the applicable contract law and varies by consequence from jurisdiction. Therefore open content licenses are often localized and can sometimes contain contractual obligations, which can result in the same doubts concerning enforceability associated with restrictive licenses.[136]

In the context of big data one of the main problems is the legal interoperability of licenses. This question is raised when datasets are combined from several sources under different licenses. Different jurisdictions with their own legal framework for intellectual property rights and for contract law can pose barriers if these frameworks are not adequately harmonised. Licenses can be used in an effort to create clear conditions of use across borders, but can also create themselves new enclosures and barriers. This is true even for the open content licenses meant to open up data sources. Also in this case the question remains how to combine data released under different licenses. When a dataset is built up through a series of adaptations and re-licensing other questions can arise, like who is liable when somewhere in the chain other rights were violated (e.g. copyrighted material was included). Open licenses with specific requirements concerning attribution or re-licensing can themselves create barriers because of lack of legal interoperability, notwithstanding their open character.

To conclude, also contract law poses problems for big data. Legal fragmentation leads to a lack of legal interoperability of licenses, and therefore to barriers for combining and using data sources. Important is that this becomes a problem for big data due to the over-protection by the IPR framework, which creates the need for open content licenses. But it is in itself also a problem for big data players who want to regulate the access to their data

[134] De Filippi, Primavera. *Copyright Law in the Digital Environment: Private Ordering and the regulation of digital works*, pp. 58-63; Bix, Brian, and Jane K. Winn, Diverging perspectives on electronic contracting in the US and the EU, *Cleveland State Law Review* vol. 54:175, 2006, 176-181.
[135] Ibid.
[136] De Filippi, Primavera. *Copyright Law in the Digital Environment: Private Ordering and the regulation of digital works*, pp. 79-82.

3.3 PRIVACY AND DATA PROTECTION

When the processing of data involves personal data the data protection framework is triggered into application. In that case the processing of personal data is constrained by the criteria and principles provided by the data protection framework and the data subject, whose personal data is processed, becomes a stakeholder in the processing with legally recognized rights. The current European data protection framework consists of the data protection directive 95/46/EC concerning processing of personal data by private and public authorities in the member states. This directive is supplemented with the E-Privacy directive 2002/58/EC concerning electronic communications. Processing by the EU institutions is regulated by regulation 45/2001. These instruments are only applicable for what were first-pillar matters, and not when the processing occurs by public authorities as part of their activities concerning police, justice or external affairs. Today the general data protection-framework provided by directive 95/46/EC is under revision. The Commission proposed a new General Data Protection Regulation (GDPR)[137] and a new directive for the processing of personal data by competent authorities in criminal matters[138]. We will focus our discussion here on the directive 95/46/EC and the GDPR, as these concern the BYTE case studies, but the problems raised concerning the application of data protection on big data in general also apply for the directive regulating data protection in criminal matters.

This data protection framework, as well as the fair information principles on which several US privacy laws are based, have received heavy criticism from industry and a range of scholars for not being suited for big data. These critics consider it being an obstacle for technical development and the scientific and economic advantages a wider implementation of big data can bring[139], or consider it broken and not effective any more to protect privacy in the age of big data[140]. Criticism has been levelled at the notions of personal data versus anonymous data, principles like purpose limitation, data minimization, and consent as base for legitimate processing of personal data. On the other hand, a range of scholars as well as the WP29 defend the application of data protection framework in the big data context and refuse to see enough ground in the fruits of progress arguments in terms of economy, security or science to lower the protection of privacy given by the data protection framework. Notwithstanding this difference in vision concerning the value of big data processing and the consequences this should have for the data protection framework, it is clear that a tension

[137] European Commission, "COM (2012) 11: Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)", *Official Journal of the European Union*, C 102, 5 April 2012, http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:52012PC0011
The legislative procedure and linked documents can be followed on http://eur-lex.europa.eu/procedure/EN/201286

[138] European Commission, "COM/2012/010: Proposal for a Directive of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and the free movement of such data".

[139] Tene, Omer and Jules Polonetsky, Big Data for All: Privacy and User Control in the Age of Analytics, 11 *Nw. J. Tech. & Intell. Prop.* 239 (2013); Ira S. Rubinstein, Big Data: The End of Privacy or a New Beginning?, NYU School of Law, Public Law Research Paper No. 12-56; Lokke Moerel, Big Data Protection: How to Make the Draft EU Regulation on Data Protection Future Proof, Tilburg University, 2014.

[140] Ohm, Paul, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, 57 *UCLA Law Review* 1701 (2010), 1701-1777; Schwartz, Paul M. and Solove, Daniel J., The PII Problem: Privacy and a New Concept of Personally Identifiable Information, *New York University Law Review*, December 2011, 1814-1894; Alessandro Mantelero, Defining a new paradigm for data protection in the world of Big Data analytics-2014 ASE BIGDATA-SOCIALCOM-CYBERSECURITY Conference, Stanford University, May 27-31, 2014.

exist between the data protection framework and big data processing. Here we will make an inventory of the problems registered in the literature.

Data protection starts to apply when personal data is processed. Personal data is defined in directive 95/46 as "any information relating to an identified or identifiable natural person", with an identifiable person being defined as someone "who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity"[141]. Both personal data ("any information") and processing ("any operation or set of operations which is performed upon personal data") are defined very broadly, so the data protection framework is triggered whenever big data processing involves data which can be linked to identifiable persons. Recital 26 states that "to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person".[142] The GDPR includes this specification in its definition of data subject.[143] In other words, the scope of the data protection framework can evolve when new techniques concerning identification in datasets become available. Result is that the nature of certain data can change over time from not being personal data into personal data, and by consequence become subjected to the data protection framework. Changing the nature of the data in the other direction is also possible. By anonymizing the data it loses its link with an identifiable person and the processing is not subjected to the data protection framework any more. Recital 26 states that "the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable".[144] Adequate anonymization is not an easy task. Main criterion for adequate anonymization is the outcome: the data subject may no longer be identifiable by the controller or a third party. If such outcome cannot be guaranteed, the data protection framework remains applicable on the dataset. As said before, such identifiability is judged taking account of "all the means likely reasonably to be used". This implies that future developments can render anonymization techniques inadequate, like new techniques but also when new data sources become public which makes such identification easier.

The Article 29 Data Protection Working Party (WP29)[145] identified 3 risks which have to be addressed through the anonymization techniques:
- singling out: the possibility to isolate records which identify an individual in the dataset
- linkability: the possibility to link 2 or more records concerning the same data subject (in one database or in different databases). When such linking would allow by combining several datasets to relate, with high probability, personal data to a specific person, the anonymization is broken. For instance, when health data from one dataset can be linked to specific individuals by using census data from another dataset.

[141] European Parliament and the Council, Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 24.10.1995, article 2(a). A similar definition can be found in European Parliament and the Council, Regulation (EC) No. 45/2001, 18 December 2000.
[142] European Parliament and the Council, Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 24.10.1995, recital 26.
[143] European Commission, "COM (2012) 11: Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)", art 4 (1).
[144] European Parliament and the Council, Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 24.10.1995, recital 26.
[145] WP29 or Article 29 Data Protection Working Party is the advisory group, consisting of representatives of member state data protection authorities and the EDPS, and established under article 29 of directive 95/46/EC. It publishes regularly non-binding opinions.

- inference: the possibility to infer with high probability the value of an attribute based on a set of other attributes. For instance, when identification in a dataset is countered by generalisation of birth dates to year of birth, but in certain age groups only one value for a specific attribute (e.g. a medical condition) is present or is dominant, it is possible to deduce this attribute with high probability for people with that age.

In its Opinion on Anonymisation Techniques, the WP29 concluded that none of the techniques currently in use could guarantee with certainty that all 3 risks of linking personal data to identifiable persons were prevented.[146] It can only be decided on a case-by-case bases if certain anonymization techniques work adequately. This makes avoiding the application of the data protection framework through anonymization a question which needs careful technical analysis. The WP29 also warned for using pseudonymization as an anonymization technique. Pseudonymization replaces an attribute, which can be used as identifier, by another. Problem is that the granularity of the data remains the same. Pseudonymization can reduce linkability with other datasets, but the new attribute can as well be used as identifier. Leaving such identifiers in the data often allows re-identification, as was shown by statistical research.

Current anonymization techniques proved to be inadequate, till the level it gets doubted that anonymization is still technically feasible in a context of big data. An obvious measure for anonymization is to drop all direct identifiers like name, address, identification codes, etc. But this proved not enough. Sweeney showed that in the US the combination of postal code, gender and date of birth made it possible to uniquely identify 87% of the population.[147] Such data functions as quasi-identifiers. In itself the individual values of the category do not structurally identify people, but combined this data can be used for identification. These quasi-identifiers are in itself not sensitive information and often already public. But they can be used to link other sensitive data to persons. A typical example of such sensitive data is health data. Medical data gets often released as a combination of occurrences of sensitive data like diseases, genetic or behavioural information (ex. smoking) with quasi-identifiers like gender, age and region. Sweeney developed a measure for privacy with the k-anonymity model. K-anonymity holds if any set of quasi-identifiers links to at least k different occurrences in the dataset. In other words, if such an occurrence is equivalent with a person, each set of quasi-identifiers is linked to at least k persons. This proved often a too weak assumption and several extra restrictions for the attributes have been proposed, depending on what information is needed to stay in the data and which identification mechanisms have to be avoided. A range of methods and algorithms have been developed for anonymization, consisting of generalization of data, adding noise or suppression of data. Such measures are also linked to certain types of data. Statistical disclosure control was first developed for tabular data, containing the aggregate results of survey data and later also the microformat data itself. Big data makes other forms of data available, like relational data in social networks, locational data or temporal data about behaviour. The differentiation between quasi-identifiers and sensitive data often becomes impossible. Other measures of anonymity for these newer forms of widely available data are object of ongoing research. The results of this research has direct implications for the interpretation in practical use cases of legal terms like personal data, and of the liability and responsibility of data controllers. Several spectacular de-anonymization attacks on real-world datasets have cast doubt that a safe

[146] WP29, *Opinion 05/2014 on Anonymisation Techniques*, 10 April 2014, p. 23.
[147] Sweeney, Latanya, 'Simple Demographics Often Identify People Uniquely'. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

release of rich datasets with relational data is possible at all.[148] This also raises the question if the existing data protection framework has reached its limits and is still capable of dealing with the privacy problems arising from the existence of much richer datasets with big data technologies.

When the data protection framework and the rights it gives to data subjects is seen as a barrier to big data processing, then keeping this broad definition of personal data will have a strong negative effect on such processing. In the draft GDPR we notice some attempts to redraw the balance and limit the effects of the wide area of application. The European Commission included in its draft an art 10, stating that "If the data processed by a controller do not permit the controller to identify a natural person, the controller shall not be obliged to acquire additional information in order to identify the data subject for the sole purpose of complying with any provision of this Regulation". This implies that when the controller limits the possibility to identify a person with the means and data under his control, the controller is not obliged to do efforts to grant the data subject access to data hold about him. For example, if the controller holds information like IP addresses or fingerprints of digital devices, but does himself no effort to link this data to identities of persons, he can refuse to give persons access about the information held by him. This article clearly narrows identifiability to identifiability by the data controller. No consent is needed and theoretically the data controller can even publish this data allowing third persons to do the identification by themselves. The European Parliament went further on this approach by adding a new category of 'pseudonymous data', defined as "personal data that cannot be attributed to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject to technical and organisational measures to ensure non-attribution". Pseudonymous data remains personal data and within the remit of the GDPR. As such it softens the implications of article 10 in the Commission draft. But the use of pseudonymous data also implies a weakening of the obligations of the data controllers by the inclusion of pseudonymous data in article 10. Result is that the access and other rights of data subject can be denied in case of pseudonymous data.

This addition of pseudonymous data is contested. As shown by the WP29 pseudonymization is not an adequate anonymization technique. On the other hand, the definition of pseudonymous data adds some extra limitations to the technique of pseudonymization and demands measures to ensure non-attribution. It can be questioned what the difference with anonymization still is. The triggering point for the data protection framework seems to move from the technical feasibility of identification, which looks at what techniques are in general available, to the effort to actually prevent identification, which includes also organisational measures to prevent the use of certain available techniques for identification. To get a more definitive view on the reach of the data protection framework, we will have to wait for the end of the legislative process. But it is clear that this legal fine-tuning will have a large impact on the control data subjects have over the processing of their personal data and on the actual big data processing involving personal data.

These contested attempts for legal fine-tuning are examples of a more general plea, often found at the critics of the current data protection framework, to move the attention of data

[148] Arvind Narayanan, Vitaly Shmatikov, 'Robust de-anonymization of large sparse datasets.(How to Break Anonymity of the Netflix Prize Dataset)', 5-2-2008, arXiv:cs/0610105v2; Arvind Narayanan, Elaine Shi, Benjamin I. P. Rubinstein, 'Link Prediction by De-anonymization: How We Won the Kaggle Social Network Challenge', 22-2-2011, arXiv:1102.4374v1

protection from data collection to the more risk-based approach based on the actual use of personal data. These proposals involve a scaled approach through which the application of data protection principles gets modulated.[149] The WP29 has reacted to this plea in its 'Statement of the WP29 on the role of a risk-based approach in data protection legal frameworks'[150] and other recent recommendations. In its statement it points to the risk-based elements present in the data protection framework, while it also makes re-interpretations of data protection principles like purpose limitation which are more compatible with this approach.

Till now we looked mostly at how the basic definitions triggered the application of data protection. Now we will look at the principles and rules to which data processing gets subjected. The directive 95/46 puts forward data protection principles to which any processing of personal data has to conform, like:
- legitimacy: all data processing has to have a legal base. The directive lists several criteria based on which the processing can happen legitimately. Main ground is the consent of the data subject. Much stricter criteria apply for a specific set of sensitive data, that is data concerning "racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and ... health or sex life".[151]
- finality: all data collection and processing has to be done for "specified, explicit and legitimate purposes". This links data collection and processing to specific purposes, specified at the moment of collection or earlier (purpose limitation). Further processing for other purposes is not allowed "in a way incompatible" with the original purposes. The WP29 points to this nuance to clarify that not all further processing is forbidden, but only incompatible use.[152] The directive also allows further processing for historical, statistical or scientific purposes.
- proportionality and relevance: the personal data must be "adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed". Only personal data that is useful and only the minimal amount needed may be processed (data minimization). A consequence is that, when the data is not necessary any more, it must be erased or kept in a form which does not allow identification. The right to be forgotten follows logically from these principles.
- accuracy : the data must be correct and up-to-date. Inaccurate data has to be corrected or erased.
- transparency : What happens to his personal data has to be transparent to the data subject. This implies that the data subject has the right to get from the controller information on the identity of the controller, the purposes of and the logic behind the processing and who receives the data.
- data subject participation and control : The data subject has the right to access the data, to object to certain processing of personal data or to demand the rectification, erasure or blocking.

---

[149] Tene, Omer and Jules Polonetsky, Big Data for All: Privacy and User Control in the Age of Analytics, 11 *Nw. J. Tech. & Intell. Prop.* 239 (2013), 258-259; Paul Ohm, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, 57 *UCLA Law Review* 1701 (2010), 1759-1777; Schwartz, Paul M. and Solove, Daniel J., The PII Problem: Privacy and a New Concept of Personally Identifiable Information, *New York University Law Review*, December 2011, 1879-1894.
[150] WP29, *Statement of the WP29 on the role of a risk-based approach in data protection legal frameworks*, 30 May 2014.
[151] European Parliament and the Council, Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 24.10.1995, article 8(1).
[152] WP29, *Opinion 03/2013 on purpose limitation*, 2 April 2013.

- data security: the data has to be kept secure to avoid unauthorised or illegitimate access, transfer or processing.

Several of these principles pose major problems in the context of big data. Most criticised are the purpose limitation and data minimization principles. Big data processing is considered "antithetical to data minimization"[153]. The value of big data is seen in the exploration of large datasets for unknown correlations, leading to new, unforeseen secondary uses. Data is collected and stored for longer periods of time than the first intended use, and combined with other datasets and re-used for new purposes. This all runs counter to purpose limitation. The big data model also pushes to broad data collection and a proliferation of sensors or data items registered, in other words data maximization rather than data minimization. The clash between the data protection model and the big data model can be seen in several high-profile court cases, like the decision of the European Court of Justice (ECJ) on the right to forget-case against Google[154] and its decision on data retention[155]. In both cases the ECJ made a clear application of the proportionality principle and the data minimization logic. The WP29 has answered to the criticisms with its Opinion 03/2013 on purpose limitation. It made clear that only incompatible uses with the stated purposes are forbidden, giving some leeway in considering further use. It also clarified the compatibility test for these uses and opened it up towards a substantive assessment taking in consideration elements like the reasonable expectations of the data subject. Purpose limitation in this way gets a bit softened into safeguarding contextual integrity, which allows more distance from the explicit articulations of specific purposes.[156] The WP29 explicitly treats the implication of purpose limitation for big data. While acknowledging the merits of big data, it also points to the privacy dangers. Concerning further use of personal data for analytics it points to the compatibility test it has drafted. Important in this assessment is the aim of this analytics: detecting trends and correlations, or analysing personal preferences, behaviour and attitudes of individuals in order to inform decision making with regard to those individuals. In the first scenario the WP29 points to the use of functional separation as an element which can influence the assessment. Functional separation means that data used for statistical purposes or other research purposes should be separated from data used for decision making concerning individuals. The further use for statistical or research purposes could be compatible. Concerning the second scenario, further use of data for decision making concerning individuals, the WP29 requires an 'opt-in' consent and access to the decisional criteria, or the profiles and the logic behind them, in order to make that consent an informed one.[157] With this recommendation the WP29 shows that the data protection framework does not make big data processing impossible. Its compatibility test for further use moves the application of the data protection principles in the direction of linking protection with control of the use of personal data instead of the collection, but without relinquishing the principles themselves. The WP29 also advises to amend some articles in the draft GDPR, which try to create more space for further use, but which in its view hollow out the basic principles too much. Again the tension between purpose limitation and big data processing is clear, while several options to deal with it are available. Important in this discussion is that the data protection framework is not just a form of consumer protection, but that purpose limitation and related data protection principles are

[153] Tene, Omer and Jules Polonetsky, Big Data for All: Privacy and User Control in the Age of Analytics, 259.
[154] EUCJ, *Google Spain SL, Google Inc. v Agencia Española de Protección de Datos (AEPD), Mario Costeja González*, C-131/12, 13 May 2014.
[155] EUCJ, *Digital Rights Ireland and Seitlinger and Others*, Joined Cases C-293/12 and C-594/12, 8 April 2014.
[156] Mireille Hildebrandt, Slaves to Big Data. Or Are We?, *IDP. Revista de Internet, Derecho y Ciencia Política*, (17), November 2013, 37-38.
[157] WP29, *Opinion 03/2013 on purpose limitation, Annex 2: Big data and open data*, 2 April 2013

clearly linked with European human rights law, which implies that every interference with a protected right has to be justified. Such interference is only allowed when it has a legal base, is for one of the purposes defined in human rights law and is 'necessary in a democratic society'. The specification of a purpose is needed to assess the necessity and the legitimacy of the interference.[158] Each effort to adapt data protection in order to enlarge the space for big data processing needs to be compatible with these basic constraints from human rights law.

The clash between the data protection principles and big data can be seen in several high-profile court cases, like the decision of the European Court of Justice (ECJ) on the right to be forgotten-case against Google[159] and its decision on data retention[160]. We will look into the right to be forgotten-case in more detail, as it is a clear example of the application of the data protection principles. The right to be forgotten has raised a lot of discussion, especially since it was included as a specific right in article 17 of the draft GDPR. It has been presented as an absolutely new right and as the "biggest threat to free speech on the Internet in the coming decade".[161] Such exaggerations forget that it is in fact an application of existing data protection principles and the right to request erasure of data is already present in article 12(b) of the Data Protection-directive 95/46. Further, it does not give people the right to erase their whole past at will. In fact, it entails no more than the right to request a re-evaluation of a specific act of processing according to the data protection principles and, when this processing proves not to comply any more with these principles, the erasure of the data.

The decision of the ECJ in the Google-case shows that such a limited right to be forgotten, or rather a right to demand the erasure of personal data, is already enshrined in the existing data protection directive. A Spanish citizen requested the removal from the Google search engine results of 2 links to articles concerning legal proceedings against him for social security debts more than 10 years earlier. The Spanish DPA had refused the removal of the original publications, as those resulted from a legal order, but had approved the request to remove the links. Google started legal proceedings against this decision, during which a ruling from the ECJ was requested. Several questions were raised, but relevant here is the actual evaluation of the removal request based on the data protection principles. The request is an application by the data subject of its right under art. 12(b) to request erasure of personal data "which does not comply with the provisions of this Directive". The Court points out that, although article 12(b) mentions "in particular because of the incomplete or inaccurate nature of the data" such non-compliance is not restricted to being incomplete or inaccurate but can also entail a non-compliance with the other quality criteria in article 6. It further states that initially lawful processing can become non-compliant after a lapse of time when the processing of the data is no longer necessary for the initial purposes. Especially when it is "inadequate, irrelevant or no longer relevant, or excessive in relation to those purposes and in the light of the time that has elapsed". When the inclusion of a link appears to be such after an evaluation requested by a data subject, it must be erased.

---

[158] WP29, *Opinion 03/2013 on purpose limitation*, 7; Fanny Coudert, Jos Dumortier and Frank Verbruggen, Applying the Purpose Specification Principle in the Age of "Big Data": The Example of Integrated Video Surveillance Platforms in France, ICRI Working Paper 6/2012, Interdisciplinary Centre for Law and ICT, K.U.Leuven; see also WP29, *Statement of the WP29 on the role of a risk-based approach in data protection legal frameworks*, 30 May 2014.

[159] EUCJ, *Google Spain SL, Google Inc. v Agencia Española de Protección de Datos (AEPD), Mario Costeja González*, C-131/12, 13 May 2014.

[160] EUCJ, *Digital Rights Ireland and Seitlinger and Others*, Joined Cases C-293/12 and C-594/12, 8 April 2014.

[161] Rosen, Jeffrey, "The Right to Be Forgotten", 64 *Stan. L. Rev. Online* 88, 13 February 2012.

With this reasoning the ECJ grounds a right to be forgotten in the data protection principles, limited to a right of erasure when the specific acts of processing are not compliant any more. That means that such obligation to erase data can only be established after a case-by-case evaluation. The ECJ gave further instructions on how such evaluation has to be done, in particular when the initial processing was based upon article 7(f) and considered "necessary for the purposes of the legitimate interests pursued by the controller or by the third party or parties to whom the data are disclosed". These interests can be "overridden by the interests for fundamental rights and freedoms of the data subject" and therefore have to be balanced. In this case it concluded that, 16 years after the facts, the general public had no preponderant interest anymore and that the right to erasure should be granted.

This case gives a clear application of the data protection principles and the balancing involved and is therefore no exceptional innovation. But it also shows how difficult the application of the data protection principles for big data operators can be. Big data can lead to big administration, in order to deal with all the requests. Google did set up a mechanism to deal with such requests and received after a year more than 280,000 requests and had to evaluate about 1 million links[162]. In other words, data protection can engender large transaction costs for big data operations. On the other hand, the response by Google to the verdict shows that compliance is not impossible. Similar mechanisms exist to remove links pointing to copyright-violating material and so on.

Using consent as ground for legitimate processing receives similar criticisms. It is linked to the principles of purpose limitation and finality, as new uses of personal data would need new consent. This is considered a flawed model leading to unnecessary burdens for both companies and data subjects.[163] A strict application would need an explicit opt-in accompanied with a specific notice for each new use. A big data-company as Google has struggled with opt-in obligations in several contexts, and defends itself that it offers an opt-out. Consent is not the only legitimation ground and these grounds could be enlarged. This would be in line with the rationale behind the plea to shift the focus of data protection from data collection to the use of data. It is also returning to the earlier model of consumer protection, from which the existing consent-based model took distance. The implementation of such shift can of course widely differ depending on how the balancing between values is conceived. Alessandro Mantelero points to earlier experiences in the mainframe era where a clear imbalance in knowledge existed between individuals and governmental or corporate mainframe operators at which a concentration of information took place. The first data protection models from this age focussed on providing a counter-control through transparency and independent control authorities. Later ICT became more widespread and accessible, diminishing the power imbalance. In this period the consent model emerged, which supposes the capacity of consumers at informational self-determination. Nowadays we see a new power imbalance and concentration of information at big data operators who control a wide range of sensors in their capacity of service providers. This concentration of information creates also a new opacity of big data processing and diminishes the capacity of informational self-determination. Therefore Mantelero also proposes to move away from the consent model,

162 http://www.google.com/transparencyreport/removals/europeprivacy/?hl=en-US, consulted on 16 July 2015.

163 Tene, Omer and Jules Polonetsky, Big Data for All: Privacy and User Control in the Age of Analytics, 260-261.

although limited to big data applications, but with counterbalancing it with independent public authorities.[164]

Another question is how the rights of data subjects concerning transparency, access and rectification are implemented in the big data context. It is possible to allow easy access to a data source through open licenses and as such create an open environment for data flows and combination of data sources. But once data sources contain personal data, new barriers arise. The first question is if the new uses are covered by the original consent, as discussed above. Even when this is the case, the following question is how to guarantee transparency, access and rectification. In a lot of cases the personal data will be anonymized and used to build profiles. From that moment the data protection framework does not apply any more. The application of the profile to see if a certain data subject fits involves its personal data and triggers the application of data protection again. This makes presume that clear delineations can be made where data protection applies and where access is granted. But when personal data is held more permanent in order to regularly rebuild profiles or other use cases which prevent the anonymization of personal data, the implementation of these rights becomes much more complicated.

To conclude, big data processing poses major problems for the data protection framework. Basic concepts, like the distinction between personal data and anonymous data, the data protection principles like purpose limitation and data minimization, and consent become difficult to sustain with big data processing. Also the rights of the data subject concerning transparency and access become more difficult to implement. If this means that the data protection framework is unsustainable in a context of big data processing, or whether this can be solved with creative interpretation and application of the existing concepts or adaptations remains an open discussion. In this context we presented the debate on the 'risk-based' approach to data protection and indicated how this debate influences the positions of official actors, like the WP29, the European Commission and the European Parliament.

In this section we limited us to the challenges big data poses to the core concepts in the data protection framework. More problems can be uncovered concerning the practical implementation of this framework. In the next chapters we will raise them in a more general context, as these problems are not limited to the data protection framework.

3.4 DUE PROCESS[165]

Big data processing can be used to inform decisions about people or even as part of automated decision making. Possible application areas are in marketing and targeted advertising, insurance, credit lending and even security-related activities. This opens a wide area of

[164] Alessandro Mantelero, Defining a new paradigm for data protection in the world of Big Data analytics-2014 ASE BIGDATA-SOCIALCOM-CYBERSECURITY Conference, Stanford University, May 27-31, 2014.
[165] Based on Hildebrandt, Mireille and Serge Gutwirth (editors), *Profiling the European Citizen. Cross-Disciplinary Perspectives*, Springer Science + Business Media B.V, 2008; Citron, Danielle Keats & Frank Pasquale, The Scored Society: Due Process for Automated Predictions, *Washington Law Review*, vol.89, no° 1 , 2014; Citron, Danielle Keats, Technological Due Process, Center for Information Technology Policy, Princeton, New Jersey, April 2009; Bart Custers, Toon Calders, Bart Schermer, Tal Zarsky (eds.), *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, Springer, 2013; Savin, Andrej, "Profiling and Automated Decision Making in the Present and New EU Data Protection Frameworks", http://openarchive.cbs.dk/bitstream/handle/10398/8914/Savin.pdf; Bygrave, Lee, "Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling", Computer Law and Security Reporter 7(2000)67-76; Tal Zarsky, "Transparent Predictions", *ILL. L. REV.*, 2013, 1503-1512.

problems, which are partially dealt with by the data protection framework but also raise issues covered by non-discrimination law, consumer protection, etc. Automated systems can have the aura of more objective decision making by eliminating human biases. However, such biases can remain present in the data, the models and the decision making criteria. And certain human 'biases', like avoiding discrimination or using other fairness criteria, we prefer to remain included. Completely automated decision making without any human intervention to check and counter-balance inputs and results is therefore generally rejected in the literature. The question remains how to include procedural safeguards and due process requirements in decision making based on big data processing. In the literature calling for technological due process we find recommendations to restore transparency of the decision making and involvement of the concerned persons by giving them the possibility to correct data and to object to decisions and to grounds on which they are based. Such transparency of decision making should also allow to check for compliance with anti-discrimination law, consumer law, etc.

A typical profiling process involves the following steps:
a) collection of data about individuals
b) building of profiles
c) applying the profiles to individuals and scoring them
d) dissemination of the scores to decision makers
e) use of the scores in decisions
Step a and b involve a large amount of persons, including persons which are not subjected to the decision making. In step a the requirements of the data protection legislation have to be followed, assuming that the profiles itself are built on anonymized data. Step c to e concern the person subjected to a decision based on the profiling. Again data protection legislation is mostly involved, but especially in step e we have to look at wider legal requirements like non-discrimination legislation.

Transparency is the key instrument on which procedural safeguards can be built. Without transparency or another possibility to get knowledge about the decision making, any possibility to object to discriminatory decisions or other irregularities remains hollow. The data protection directive gives the data subject the right to access personal data held by the data controller, as well as to obtain "knowledge of the logic involved in any automatic processing of data concerning him at least in the case of the automated decisions referred to in Article 15 (1)".[166] Such an automated decision is a decision "which produces legal effects concerning him or significantly affects him and which is based solely on automated processing of data intended to evaluate certain personal aspects relating to him, such as his performance at work, creditworthiness, reliability, conduct, etc.".[167] This definition is quite restrictive. From the moment a human evaluates the results of the automated processing before confirming the decision, it is not subject any more to this rule concerning the logic. It is further restricted by the reasoning that "this right must not adversely affect trade secrets or intellectual property and in particular the copyright protecting the software".[168] This allows data controllers to give a basic description of the logic, without going into much detail about criteria and weighting. But if such description is enough to be able to determine compliance

[166] European Parliament and the Council, Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 24.10.1995, art. 12(a).
[167] European Parliament and the Council, Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 24.10.1995, art. 15(1).
[168] European Parliament and the Council, Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 24.10.1995, recital 41.

with rules concerning decision making and treatment of persons in non-discrimination or consumer law can be doubted. This shows the different aims of the legislation involved. Data protection aims at informational self-determination, or assuring that individuals can decide what happens with information concerning them. The limited transparency concerning the logic can in this case be sufficient. Other legislation forbid certain treatment of people and the decisions to do so, like treatment different based on sex, race, etc. In order to check for these biases and to test if such decision making leads to forbidden discrimination, a more profound transparency is needed. Such testing, or more generally the introduction of procedural safeguards, can also be done through other actors, like supervisory authorities, and means, like auditing and impact assessments. But it requires opening up such actors and instruments from one field to the requirements of another. Concerns of discrimination have to checked substantially in earlier steps than step e mentioned above and these steps have to be open for testing and auditing in order to assure technological due process.

This problematic of due process in automated decision making is wider than big data, while big data processing can further complicate these requirements of technological due process. It can involve the use of data sources of which the relevance and accuracy is unclear. Data mining is often used to build profiles based on a wide variety of data sources and of which the relevance is uncovered by the algorithm used. Which information is exactly used and how strong it defines the end result can be very opaque. The resulting level of accuracy is often rather low and human discretion plays an important role, as acceptable error rates are set by the user of the data mining method. Certain data mining methods also merge the profile building and application steps mentioned above. This is often the case when unsupervised learning and clustering methods are used. All this even augments the need for transparency and the possibility for auditing and testing.

Due process is more than transparency and concerns the actual capacity of a person to being heard (and not just observed), to question decisions and to object to the use of certain criteria, models or data. But transparency is a key building block without which such capacity remains hollow.

To conclude, big data processing changes decision making and the construction of facts on which these decisions are based. This raises the question of how to sustain the autonomy and capability to act of persons. An important approach for this is looking how due process can be ensured in the context of big data processing. A key building block in such due process mechanisms is a larger transparency, which should also allow a larger insight in the logic behind the decision making. This should not only be adequate for the data subject's informational self-determination, but also allow for auditing and testing in order to evaluate decision making in the context of requirements of other legislation, like non-discrimination law or consumer protection. Instruments like impact assessments have to include requirements from other legal frameworks and not be limited to data protection

3.5 LIABILITY

Accountability involves issues of liability, who is responsible for which fault, and issues of jurisdiction, that is which laws apply and which courts can deal with the problem. Both these aspects of accountability become more complicated with big data processing and cloud computing as underlying infrastructure. Again we notice that the legal frameworks were conceptualized in another technological environment and with other use cases as reference. In this section we consider liability, in the next one we deal with issues of jurisdiction.

Big data processing relies on distributed computing and cloud computing is the enabling technology for that. This new technological environment can lead to very complicated set ups involving a lot of actors. Cloud computing services enable the shift of computing resources and activities from an in-house activity to services provided over the internet. Although they exists since the beginning of the internet (web hosting services are an early example), they have only recently became available for a wider range of purposes. This shift turns the internet from a communication tool between distinct companies and customers into an environment in which also internal business processes are taking place (which would before happen on the internal intranet of a company). Cloud computing allows a more flexible use of resources based on need and a further outsourcing of activities. Underlying mechanism is virtualization, through which activities can be abstracted from the underlying infrastructure or resources, which get allocated from a shared pool over the Internet. This virtualization is possible on many layers and can lead to a very complicated architecture with several partners involved. Cloud computing is offered over a spectrum of services from low-level Infrastructure-as-a-Service (IaaS), over Platform-as-a-Service (PaaS) where tools are provided to build custom applications, to Software-as-a-Service (SaaS) where end-user applications are offered. Such end-user applications can be built on platforms and infrastructures of other cloud providers and involve other SaaS layers as well. For instance, Dropbox offers a SaaS but functions on top of Amazon's EC2 IaaS. Result is a chain of outsourced activities of which the end user often has no idea.

In contrast to that, the legal frameworks organise accountability with more simple use cases as reference, which now often get combined and blurred in a complicated architecture. Big data processing and cloud computing reflects the technical convergence and reordering of services which are governed by different acts of community legislation. Here we will focus on the processing of personal data to illustrate the issue, but a similar account can be given on other issues.[169]

The Data protection directive 95/46 divides responsibility between data controllers and data processors. The data controller is the person or entity that determines how and for what purposes personal data gets processed. He can outsource such processing to others. These persons or entities, called data processors, are considered to work on behalf of the controller and under his instructions. The Data protection directive clearly forbids the data processor to process personal data outside of the instructions given by the data controller or obligations provided by law.[170] It also requires that such outsourcing is governed by a written contract or legal act which includes this and other data security requirements.[171] The data controller carries the main liability for processing personal data in violation of the data protection law. When he can prove he is not responsible for the damaging event, he can be exempted.[172] This would imply he has taken all necessary and reasonable measures to prevent data breach or misuse of data, otherwise he is responsible as result of his negligence. Important in this context is that the data security obligations in the directive are not just addressed to the data controller, but also to data processors. The data processor is responsible for the data security

[169] Verbiest, Thibault, Gerald Spindler, Giovanni Maria Riccio, Aurélie Van der Perre, *Study on liability of internet intermediaries*, 12 November 2007, p. 13.
[170] European Parliament and the Council, Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 24.10.1995, art. 16.
[171] European Parliament and the Council, Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 24.10.1995, art. 17(3).
[172] European Parliament and the Council, Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 24.10.1995, art. 23.

aspects linked to his task and the liability for unlawful processing can even shift totally to the data processor if the data controller has done all what could be expected to assure secure processing (like giving adequate instructions and checking if the data processor could be considered up to his task and had all the necessary arrangements in place to adequately deal with the data).[173] This legal set up is clearly made with ordinary outsourcing situations in mind, where a person or company is hired to do administrative, marketing or other tasks involving personal data and where the chain of outsourcing is short. Further, it also implies that all subcontractors involved know they are dealing with personal data and have to treat it accordingly. It can be questioned if this legal set up is still followed in complicated cloud computing set ups. For example, if the obligation to point out the responsibilities of the data processor in a written contract is indeed applied along the whole chain of subcontracting. It would be responsibility of the data controller to ensure this, but it is doubtful this is even possible. Data security is important for the processing of non-personal data as well, but in the context of personal data it is governed by specific legal rules. It is doubtful if all providers of cloud services know what is personal data and what not, and are by consequence aware when they become data processors as defined in the data protection laws. In that sense they more resemble like the intermediary service providers in the Directive 2000/31/EC on electronic commerce, but the data protection framework does not include such a position and the electronic commerce directive excludes the aspects covered by the data protection directives from its application.

The Directive 2000/31/EC on electronic commerce has a different set up of liability. It includes a limitation of liability for intermediary service providers (like ISPs, e-mail services, web hosting, …) giving access to communication networks, transmitting communications or providing storage. It excludes these providers from liability for information transmitted or stored, when they have no active part in the transmission apart from 'mere conduit' or have no knowledge about the content or its illegality and when they acts expeditiously to remove or to disable access to the information when they obtain such knowledge or awareness.[174] This supposes service providers which treat data or information as black boxes of which they do not have more knowledge. Their exemption ends when they receive such knowledge (e.g. by a take down-notice), but they have no active duty to control the legality of the information stored or transmitted. And as mentioned, this directive excludes data protection from its application. These service providers remain liable as processor of personal data, although in most cases the data security and confidentiality rules of Directive 2002/58/EC on privacy and electronic communications will apply. This directive implements data protection "in connection with the provision of publicly available electronic communications services in public communications networks".[175] The rules in this directive are better adapted to guard the black box character and specify for different types of data connected with the communication service how they have to be treated.

To which service providers the limitation of liability applies can be different depending on the national implementations of the electronic commerce directive. For example, some of these national implementations deal also with search engines or linking.[176] It could be argued that

[173] European Parliament and the Council, Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 24.10.1995, art. 17.
[174] European Parliament and the Council, Directive 2000/31/EC on electronic commerce, 8 June 2000, art. 12-14.
[175] European Parliament and the Council, Directive 2002/58/EC on privacy and electronic communications, 12 July 2002, art. 3(1).
[176] Verbiest, Thibault, Gerald Spindler, Giovanni Maria Riccio, Aurélie Van der Perre, *Study on liability of internet intermediaries*, 12 November 2007, p. 84-106.

this limitation of liability also applies to several cloud computing services. Again, we see that this limitation of liability is made with certain use cases as reference (ISPs, e-mail-services, website hosting) but that new developments blur the area of application. And these new development like cloud computing fall outside the more strictly defined boundaries of the e-privacy directive, which brings them in the general regime of directive 95/46.

Although the delineation between these directives is clear from a legal point of view, it is much less so in practice. The technical convergence and reordering of services which come along with cloud computing and big data processing can make it difficult to discern where we shift from the binary data controller-data processor set up into the more black box-oriented arrangement of intermediaries in the e-commerce and e-privacy directive. Again the question can be raised if the legal frameworks are not in need of clarification or updating to the more complex technical environment.

3.6 JURISDICTIONAL PROBLEMS

Big data processing can run into similar problems concerning jurisdiction, leading to a large amount of applicable laws and high costs to assure compliance with all of these. Jurisdiction defines which laws apply and which courts can deal with the problem. An extensive formulation of what elements make a legal framework applicable can make laws from several jurisdictions applicable on the same activity. Both the large-scale combining of data sources originating from several jurisdictions and the technical convergence of services mentioned in the last section can lead to an inflation of applicable frameworks.

For example, national data protection laws in the EU become applicable on processing of personal data by non-EU data controllers when this processing "makes use of equipment, automated or otherwise, situated on the territory of the said Member State, unless such equipment is used only for purposes of transit through the territory of the Community".[177] This can make EU data protection rules applicable on non-EU data controllers processing personal data from non-EU citizens but with use of a data centre or cloud services situated in an EU Member state. And this for all stages of the processing, also those outside the EU.[178] It is possible that the data controller is not aware of all applicable laws as the data controller can be unaware of some of the layers part of the complicated architecture underlying cloud computing services. EU-based data controllers are also easily confronted with the parallel application of national laws on different sets of data or even the applicability of a range of data protection laws on the same dataset.[179] Similar problems can arise due when the big data processing uses data sources from a variety of jurisdictions.

[177] European Parliament and the Council, Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 24.10.1995, art. 4(1)c.
[178] Korff, Douwe, *New Challenges to Data Protection Study - Working Paper No. 2: Data Protection Laws in the EU: The Difficulties in Meeting the Challenges Posed by Global Social and Technical Developments*, European Commission DG Justice, Freedom and Security Report, 15 January 2010, pp. 29-38; Hon, W. Kuan, Julia Hörnle and Christopher Millard, Data Protection Jurisdiction and Cloud Computing – When are Cloud Users and Providers Subject to EU Data Protection Law? The Cloud of Unknowing, Part 3, International Review of Law, Computers & Technology, Vol. 26, No. 2-3, 2012; Queen Mary School of Law Legal Studies Research Paper No. 84/2011, 9 February 2012; Article 29 Data Protection Working Party, *Opinion 8/2010 on applicable law*, WP179, 16 December 2010, pp. 20-25.
[179] Korff, Douwe, *New Challenges to Data Protection Study - Working Paper No. 2: Data Protection Laws in the EU: The Difficulties in Meeting the Challenges Posed by Global Social and Technical Developments*, European Commission DG Justice, Freedom and Security Report, 15 January 2010, p. 26.

The draft GDPR tries to address the problem by changing the linking through the location of equipment into a link of the processing activities with the offering of goods or services to data subjects in the EU or the monitoring of their behaviour.[180] This already limits the extra-territorial applicability due to the mere use of a cloud service in the EU. Instead the processing has to concern data subjects in the EU. The problem of the applicability of several laws of EU Member states will be solved with the GDPR being a regulation which does not require implementation in national law and therefore being directly the sole applicable law in the whole of the EU. Problems due to the use of data sources originating from many jurisdictions will also be diminished with this harmonization within the EU by the GDPR, but remain when also sources from outside the EU are used. It can be further alleviated by using the instruments developed to simplify the international transfer of personal data, like the 'Binding corporate rules' (BCR).

In general legislators are tempted to include light application mechanisms in their legal frameworks when the phenomena to regulate start to escape national boundaries. This is particularly the case for internet-related activities. Fear is that these phenomena would otherwise escape regulation. But the result of this approach can become an inflation of applicable laws which start to have extra-territorial reach. The better approach is harmonization or at least the introduction of flexible harmonizing instruments into the legislative frameworks. This is also the case in the data economy, where legal interoperability of data sources and data processing activities and infrastructure is a key enabler.

We can conclude that the virtualization processes underlying big data processing challenge the functioning of basic legal mechanisms. In this section we dealt with jurisdiction, where we see that light linkage mechanisms are used to trigger laws into application. This assures that no activity escapes legal regulation, but also leads to an inflation of applicable laws. Harmonization mechanisms are the best solution to create a stable legal environment for big data processing.

3.7 CONCLUSION

In this chapter we have made an inventory of the challenges that big data processing poses to legal frameworks regulating the processing of data, and vice versa of the hurdles these legal frameworks pose to big data processing. As raised in the introduction, the language of externalities is a bit difficult to integrate with legal issues. In fact the more important question is if these legal frameworks can be made compatible with big data processing without or with small changes. Or that a major overhaul is needed, and which balance to strike in that case.

The protection by copyright and sui generis database rights of data sources clearly limits big data processing. It sets up isolated data sources, and making them available involves high transaction costs due to obtaining the necessary licenses for each source. Restricting the protection would give space for data flows and combination of data sets, as well as for new uses like data mining. This comes down to redrawing the balance on which the IPR framework is based in order to let it better achieve its objectives. The legal fragmentation also has a negative impact on legal interoperability of data sources. Tools exists to enhance the legal interoperability, like open content licenses, but more legal harmonization is a more fundamental solution.

[180] European Commission, "COM (2012) 11: Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)", art. 3(2).

Big data processing poses major problems for the data protection framework. Basic concepts, like the distinction between personal data and anonymous data, the data protection principles like purpose limitation and data minimization, and consent become difficult to sustain with big data processing. Also the rights of the data subject concerning transparency and access become more difficult to implement. If this means that the data protection framework is unsustainable in a context of big data processing, or whether this can be solved with creative interpretation and application of the existing concepts or adaptations remains an open discussion.

Big data processing changes decision making and the construction of facts on which these decisions are based. This raises the question of how to sustain the autonomy and capability to act of persons. An important approach for this is looking how due process can be ensured in the context of big data processing. A key building block in such due process mechanisms is a larger transparency, which should also allow a larger insight in the logic behind the decision making. This should not only be adequate for the data subject's informational self-determination, but also allow for auditing and testing in order to evaluate decision making in the context of requirements of other legislation, like non-discrimination law or consumer protection. Instruments like impact assessments have to include requirements from other legal frameworks and not be limited to data protection. Due process is more than transparency and concerns the actual capacity of a person to being heard (and not just observed), to question decisions and to object to the use of certain criteria, models or data. But transparency is a key building block without which such capacity remains hollow.

A last important issue is the impact of the virtualization processes underlying big data processing on the functioning of basic legal mechanisms. Big data processing relies on distributed computing and cloud computing is the enabling technology for that. Such cloud services can be combined and layered, leading to complicated architectures which can be opaque. Also the technological convergence of services leads to problems in the application of legal frameworks, as these often were conceptualized on distinct use cases which now get blurred. Both the opacity and the blurring of the application of legal frameworks creates difficulties for the application of liability mechanisms, and makes the need arise for clarification or updating of these legal frameworks to the more complex technical environment. Similarly this creates jurisdictional problems, including a risk for the inflation of applicable laws, which raises the need for harmonization.

As a general conclusion we see that existing legal frameworks provide impediments to big data processing and that big data processing challenges the functioning of several of these legal frameworks. A clear need for legal reform is present in order to reap the advantages of big data, while restoring the protection of values and interests where such protection is endangered.

4 SOCIAL AND ETHICAL ISSUES IN BIG DATA

Anna Donovan and Rachel Finn
Trilateral Research & Consulting

4.1 OVERVIEW

The current culture of big data and the relevant information technology practices associated with big data, including transparency, personalisation techniques, profiling, tracking, re-use, unintended secondary use, sharing, data access and open access enable the extraction of knowledge and insights from large and complex collections of digital data. However, these practices raise a number of social and ethical issues including trust, discrimination, privacy, and exploitation. The volume, variety and velocity of big data exponentially increase the positive externalities produced by big data but also increase the risk of social and ethical values being compromised through the utilisation of big data technology practices, especially as the technologies used for those practices are ethically and socially neutral. The recognition of social and ethical issues in connection with big data are important because the implementation of big data technologies and practices that are considerate of social and ethical values can support the sustainability of the big data industry. This chapter undertakes a broad-brush examination of the different social and ethical issues that are relevant to big data currently by arising in relation to the aforementioned technology practices, and which may be relevant to big data as it further develops. This discussion is useful because social and ethical issues result in positive and negative externalities for society and have implications not only for citizens but also for industry wishing to capture the benefits that flow from big data.

Big data technologies and practices heavily involve people and their personal information. The relationship between people and big data is the basis for the likelihood of social and ethical issues arising in relation to big data technologies and practices. Davis and Patterson suggest that big data technologies present social and ethical issues because they relate to personal values and how to apply such values to the production and consumption of products and services that utilise big data.[181] Hence, there is considerable overlap between issues that may be deemed "social issues" and others that may be deemed "ethical issues" in the context

[181] Davis, Kord and Doug Patterson, *Ethics of Big Data,* O'Reilly Medica, Inc., California, 2012, p.5.

of big data, although the following general distinctions can be drawn. Social issues are those that affect the fabric of society to the extent that they are harmful to society, or produce a benefit. Big data is poised as producing a number of social benefits such as those relating to health care, although there are social issues that produce negative externalities such as the practice of discrimination that may flow from data profiling techniques. On the other hand, ethical issues are issues that have an inherent right vs. wrong quality that can affect how people make decisions and lead their lives. According to Fule and Roddick, ethics refers to:

> A set of moral principles or a system of values, which guides the behaviour of individuals and organisations. It is the correct way of doing things, which is judged by society and often enforced through law (such as anti-discrimination legislation). To act ethically involves acting for the benefit of the community. It is entirely possible to act unethically yet legally.[182]

For example, big data technologies can compromise values such as privacy and thereby, erode users' trust in online services such as social networks and other free public web based services such as webmail. However, despite there being a theoretical distinction between social and ethical issues, practically speaking, a number of issues that arise in relation to big data comprise both ethical and social aspects and do not necessarily conform to just one understanding of what is meant by a social issues versus ethical issues. This is the case for a number of the issues addressed in this report such as trust, discrimination, inequality of access to data, re-use of data, unintended secondary use and sharing, exploitation, manipulation and privacy. Increasing public concern about having personal data captured, aggregated, sold, mined, re-sold and linked to other data heightens the importance of recognising these issues. However, the interrelationship between social and ethical issues in the context of big data is made more complicated by the fact that what is ethically and socially acceptable is often largely a matter of personal opinion as people impose their own code. What might be considered an ethical issue for one person may be considered a social one for another, or conversely, present no issue for the citizen at all. Although, it is likely that issues having a direct effect on a citizen's behavior, reputation or identity will arouse stronger objections on the basis of ethics or social appropriateness. As big data technologies and practices have produced a number of positive externalities such as the increase in transparency in scientific, commercial and government decision making,[183] these often overwhelm the negative externalities that arise when social and ethical values are compromised in the context of big data. Further, a number of difficulties arise in attempting to diminish negative social and ethical impacts on individuals and society because they can be subconsciously and subjectively imposed on big data technology practices such as search algorithms, as seen in the case of discrimination. Ultimately, as big data develops, big data companies can recognise the importance of taking into consideration the social and ethical issue raised in relation to technologies and practices, and potentially incorporating social and ethical values into their policies. For big data companies in the future, this mode of operation can minimise practices that compromise social standards and ethical values, and carve out a unique position to better capture the positive externalities that flow from big data.

4.2 SOCIAL AND ETHICAL ISSUES

[182] Fule, Peter, and John F. Roddick, "Detecting Privacy and Ethical Sensitivity in Data Mining Results", *Conferences in Research and Practice in Information Technology*, Australian Computer Society, Inc., Dunedin, New Zealand, Vol. 26, 2004, p.1.

[183] Royal Society, *Science as an Open Enterprise*, June 2012. http://royalsociety.org/policy/projects/science-public-enterprise/report/

Vast amounts of personal data are provided by individuals to make up big data collections. This human element of big data means that ethical and social issues are often intertwined. Such issues can either represent a social or negative externality and/ or produce positive and negative externalities for citizens and society.

4.2.1 Social issues

Social issues are those that have an impact on society. Positive externalities produced by social issues include those that are recognised as creating benefits for society, such as health or consumer-related benefits. These positive externalities are a direct result of big data information practices including transparency, profiling, tracking, re-use, unintended secondary use, sharing, data access, open access and manipulation. The Royal Society observes that there are social benefits to be derived from technology practices associated with big data:

> There are clear social benefits to be derived from big data analytics, for example in scientific and medical research. Being transparent about the purpose and impact of the analytics can also have benefits, in helping people to be confident as 'digital citizens' in a big data world.[184]

However, the same big data technology processes that produce the aforementioned benefits also compromise social values that can hinder an individual's standing or participation in real space as well as digital society. Thus, positive externalities associated with big data should not automatically take precedent without considering their potential for eroding traditional social values. Andrejevic opines, "data processing results in social problems that are "a consequence of the commercial system that we are creating to support the rapidly growing digital communication infrastructure".[185] This is particularly so because of "the increasing migration of social and economic activities on line".[186] In particular, the emergence of social media in the mid 2000s presented a number of new social issues:

> For the first time, we can follow imaginations, opinions, ideas, and feelings of hundreds of millions of people. We can see the images and the videos they create and comment on, monitor the conversations they are engaged in, read their blog posts and tweets, navigate their maps, listen to their track lists, and follow their trajectories in physical space. And we don't need to ask their permission to do this, since they themselves encourage us to do by making all this data public.[187] Social network site Facebook, for example, now counts over 900 million active participants around the world generating together more than 1500 status updates every second about their interests and whereabouts. In 2011, e-commerce platform eBay collected data on more than 100 million active users including the 6 million new goods they offered every day.[188]

---

[184] UK information Commissioner's Office, *Big Data and Data Protection,* UK ICO, UK, 2014, p.4. http://ico.org.uk/news/latest_news/2014/~/media/documents/library/Data_Protection/Practical_application/big-data-and-data-protection.pdf

[185] Andrejevic, M., "Exploitation in the Data Mine", in Fuchs, C. et al., (Eds.), Internet and Surveillance: The Challenges of Web 2.0 and Social Media, Routledge, London,, 2012, p.82.

[186] Organisation for Economic Cooperation and Development ("OECD") , "Harnessing data as a new source of growth: Big data analytics and policies", *OECD Technology Foresight Forum 2012,* Paris, 22 October 2012.

[187] Manovich, Lev, "Trending: the Promises and the Challenges of Big Social Data", *manovich.net,* 28 April 2011, p.2. http://manovich.net/content/04-projects/065-trending-the-promises-and-the-challenges-of-big-social-data/64-article-2011.pdf

[188] OECD, op.cit.,2012.

Thus, big data is important because "it changes the social contract".[189] There remains truth in the statement "technology's interaction with the social ecology is such that technical developments frequently have environmental, social, and human consequences that go far beyond the immediate purposes of the technical devices and practices themselves.[190] In fact, some observers believe that the social issues raised by data technologies and practices requires serious attention because:

> The digital communication infrastructure and the larger social problems associated with data processing cannot be solved by strengthening consumer consent or enabling individual consumers to block data processing. Instead, it requires a rethinking of the whole social system associated with big data and consumption. Žižek concurs, stating that the "proper answer to this [surveillance] threat is not to retreat into islands of privacy, but an ever stronger socialization of cyberspace.[191]

4.2.2 Ethical issues

Similarly to social issues presented above, potential ethical issues arise in relation to big data practices including transparent practices, personalisation techniques, profiling, tracking, re-use, unintended secondary use, sharing, data access and open access. Issues such as privacy, exploitation and manipulation often occur at the expense of user morality. This means there are serious issues involved in the ethics of online data collection and analysis. The relationship between big data and ethics is thought to be pushing us "[…] to consider serious ethical issues including whether certain uses of big data violate fundamental civil, social, political, and legal rights."[192] Thus, awareness of ethical issues arising in relation to big data technologies and practices produces a valuable benefit to society. By raising awareness of these issues, citizens may be more likely to act in a manner that requires big data companies to modify their practices, especially where privacy violations are concerned. Raising awareness of these ethical issues has also led to some big data companies implementing more ethically focused policies relating to the way in which they handle data. Fule and Roddick's consider, "Ethics is a loaded word […] even the word itself can imply judgment: do-this-don't-do-that kinds of directives and obligations." However, Sollie and Duell observe that "although technology is easily one of the most permeating and consequential features of modern society, surprisingly, an ethics of technology is still in its infancy".[193] Very little is understood about the ethical implications underpinning the big data phenomenon, and raises questions such as:

> Should someone be included as a part of a large aggregate of data? What if someone's 'public' blog post is taken out of context and analyzed in a way that the author never imagined? What does it mean for someone to be spotlighted or to be analyzed without

---

[189] Peters, Brad, "The Age of Big Data", *Forbes online,* December 2012. http://www.forbes.com/sites/bradpeters/2012/07/12/the-age-of-big-data/

[190] Cited in Boyd, Danah and Kate Crawford, "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon", *Information, Communication & Society,* Vol. 15:5, 2012, pp. 662-6679, p.662.

[191] Cited in Finn, Rachel and Kush Wadhwa, "The Ethics of 'Smart' Advertising and the Regulatory Initiatives in the Consumer Intelligence Industry", *Info,* 2014, p. 25.

[192] Davis, Kord and Doug Patterson, *Ethics of Big Data,* O'Reilly Medica, Inc., California, 2012, p.8.

[193] Wright, David, "A Framework for the Ethical Impact Assessment of Information Technology", *Ethics and Information Technology,* Vol. 13, No. 3, September 2011, pp.199-226, p. 203.

> knowing it? Who is responsible for making certain that individuals and communities are not hurt by the research process? What does informed consent look like?[194]

Important reasons for this 'underdevelopment' of a methodology for morally evaluating technology development are thought to be related to its complex, uncertain, dynamic, and large-scale character that seems to resist human control.[195] Nevertheless, ethical issues will continue to arise in relation to big data technologies and practices because technology is "ethically neutral"[196], but at the same deals with human social and moral codes. Further,

> Ethics is a highly personal topic and one that comes loaded with lots of polarising vocabulary such as good, bad, right and wrong. We all have personal moral codes, which naturally vary from individual to individual. The lack of common vocabulary for expressing the relationship between what we personally believe in and what we, as members of a common enterprise, plan to do with big data can create constraints on productive discussion and obstacles to finding consensus.[197]

Thus, the ethics of big data generally requires recognition so that an understanding may be developed, and a respect fostered, for the ethical values implicated by big data technologies and practices, so that such values can be incorporated into big data policies. This can ultimately promote sustainability of the big data industry. To that end, we may see the development of more commercial policies that promote ethical values.

4.2.3 Summary

Therefore, there is considerable overlap between social and ethical issues arising in relation to big data information technologies and practices. This is because of the human element of the data being collected and processed. Big data technologies that perform anti-social behaviours, such as profiling and tracking, which can result in discrimination, and practices that raise trust issues, privacy, exploitation and manipulation for example, comprise social and ethical aspects that can have negative implications for users and society. The potential growth of the big data industry can eventually be limited by a lack of recognition of how these issues affect citizens and society as a whole. This is especially the case if big data business models disregard longstanding social and ethical values, especially when they disregard results in negative externalities for citizens who are current and future data subjects. The following section examines a number of social and ethical issues in detail and determines their relevance to big data technologies and practices.

4.3 BIG DATA TECHNOLOGIES AND PRACTICES THAT RAISE SOCIAL AND ETHICAL ISSUES

4.3.1 Transparency

Transparency in big data practices is an important element of the relationship between big data subjects (those providing the data) and big data companies and other public and private organisations wishing to capture the benefits of big data. Transparency of practices can produce positive and negative implications for big data companies and users. On one hand,

---

[194] Boyd, Danah and Kate Crawford, "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon", *Information, Communication, & Society*, Vol. 15:5, p. 662-679, p.19. http://www.danah.org/papers/2012/BigData-ICS-Draft.pdf

[195] Cited in Wright, op. cit., 2011, p. 201.

[196] Davis and Patterson, op. cit., 2012, p.8.

[197] Ibid., p.9.

transparency builds user trust and in turn, promotes the disclosure of more data by trusting data subjects, whilst on the other hand, it can cause users to modify or distort their behaviour and limit the amount of data they provide or perform data sabotage. Whether transparency has positive or negative implications for big data, those implications are connected to levels of trust users hold in big data companies and organisations performing the relevant practices. Trust is a multifaceted issue, and whilst is has social ramifications and forms the basis for social interaction, it also comprises ethical considerations when considered in light of big data information technology practices that are perhaps less opaque than they could be, such as data mining and tracking.

Positive implications of transparency

The relationship between trust and transparency is ever relevant to the big data industry. Transparency in conducting big data processing such as data mining and data analytics is a precondition to public trust and confidence. A lack of transparency risks undermining support for, or interest in, a technology.[198] The UK Information Commissioner's Office (ICO) suggests transparency is the key to fostering trust by big data companies. The ICO observes:

> There is evidence that some companies are developing an approach to big data that focuses on the impact of the analytics on individuals. This approach is not concerned solely with the capabilities of the analytics or their compliance with data protection legislation, but also looks to place big data in a wider and essentially ethical context. In other words, they are asking not only "can we do this with the data?", i.e. does it meet regulatory requirements, but also "should we do this with the data?" i.e. is it what customers expect, or should expect?[199]

In this regard, the ICO proffers the example of a global company, Aimia, which operates in the business of loyalty management and runs loyalty programmes such as Nectar dealing with large volumes of customer data. Aimia's core values include transparency and trust and aim to implement this strategy to guide data usage and processing only in a way that data subjects have agreed to, and ultimately build their trust.[200] This is perhaps indicative of a new trend of big data companies achieving sustainability in their processes and activities by respecting data subjects and thereby, ensure users continue to participate by disclosing their data. It follows that the greater the level of trust, the greater the amount of data available for use. Another example of big data companies implementing transparent practices is provided by the IBM ethical framework, which recognises users wish to retain some control over their data (even where they are mandatorily required to provide it in exchange for a digital service). That sentiment prompted IBM to be more mindful of implementing more transparent practices to abate concerns relating to the exploitation of users' information. IBM have been developing an ethical framework for big data analytics which takes into account the context in which the data is collected and used, whether people will have a choice in giving their data, whether the amount of data and what will be done with the data is reasonable in terms of the application, the reliability of data, who wins the insights to be gained from that data, whether the application is fair and equitable, the consequences of processing, people's access to the data; and accountability for mistakes and unintended consequences. Thus, data companies can display their respect for users' values by operating in a transparent manner. This business insight is becoming increasingly important as commercial organisations become more aware of the trust issues consumers have in relation to the use of their data, particularly sensitive

[198] Wright, op. cit., 2011, p. 213.
[199] UK information Commissioner's Office, op. cit., 2014, p.44.
[200] Ibid., p. 47.

personal data.[201] In the absence of transparent policies, big data processes can become shrouded in suspicion, which can lead to a growing distrust amongst users of digital services. This effect of this for big data industry is great when considering distrust can lead to reluctance by individuals to freely provide their data and make uninhibited use of digital services.

Negative implications of transparency

However, transparency is not always viewed as the answer to increasing the flow of data and subsequently, prosperity for big data companies and organisations. However, transparency can also lead to negative societal outcomes, such as practices like 'data sabotage', whereby once actors realise that an institution is collecting data and looking for patterns, they can attempt to sabotage this by providing false information. What this means is that big data companies can face an ethical dilemma in deciding whether to make practices transparent that may ultimately result in less data for them, versus conducting opaque practices that potentially compromise values such as privacy, with the view that less transparent practices may lead to the collection of greater amounts of data, irrespective of the ethics associated with doing so. For example, it is ethically and legally questionable whether big data companies can secretly collect, track, mine and analyse an individual's data, but on the other hand, the overriding commercial goal of a big data company is to collect and make use of the data. In order to do that, the greater amount of data collected directly correlates with the commercial gain. Furthermore, transparency can benefit users whilst, at the same time, represent a negative commercial externality for big data companies that implement transparency as a means of gaining user trust, such as causing a chilling effect. A "chilling effect" has been identified as arising in response to the collection of personal information for the purpose of constructing profiles for personal or targeted advertising whereby individuals are discouraged from conducting Internet searches, making purchases or using specific consumer services because of personalised advertising applications.[202]

Nevertheless, as trust is a vital ingredient in the relationship between data subjects and big data companies and organisations, transparent practices should be favoured over opaque practices to build user trust and achieve increased collections of data without compromising ethical values.

4.3.2 Profiling and tracking

Profiling, particularly as a result of tracking, is a common big data information technology practice that can have positive and negative implications for users. Whilst it enables websites and advertisers to tailor their advertisements and provide a more personalised service for users, in some cases, it can cause discrimination, or amount to the exploitation of consumers.

Negative implications of profiling and tracking: discrimination, exploitation and manipulation

Discrimination has social ramifications both in real space and for users participating in the digital society. Discrimination can result from profiling techniques that reflect perceived and existing social biases. Finn and Wadhwa observe, "the sole purpose of profiling is to discriminate between different categories of consumers along the lines of gender, wealth and

[201] Ibid.
[202] Finn and Wadhwa, op. cit.. 2014, p.7.

age."[203] For example, discrimination can result from profiling where individuals are identifiable as a result of the creation of links between an individual and particular, sensitive personal characteristics, specifically in the case of personalised advertising.[204] In addition, practices such as data mining in state and security applications combined with profiles generated by the consumer intelligence industry or consumer credit reports matched to government data can reinforce existing inequalities as well as make deferential power relations opaque.[205] This can ultimately lead to discrimination and result in adverse decisions being made against individuals in areas such as finance and employment. An example of potential discrimination is evident when social data is obtained from users' social media accounts. That information enables an organisation to paint a picture of their potential customer base and impose discriminatory or prejudicial biases. In 2013, profiling with social media data was attempted by Germany's largest credit agency, Schufa.[206] This practices sparked public outcry after several news outlets reported that it was planning to scrape data from social networks to gauge a consumer's creditworthiness, and it consequently backed down from employing this technique. This specific example also represents the potential for when ethical values, such as privacy, are automatically compromised. Such profiling techniques are potentially privacy invasive and likely not what the social network user had in mind when they were updating their social network profiles. Another example of this compromising practice is evidenced when credit applications are categorically denied on the basis of Facebook likes and preferences. Credit companies in America, such as Lenddo, Neofinance and Affirm for example, believe "a person's social standing, online reputation and/or professional connections are factors that should be considered when extending credit, especially to someone with a scant or spotty credit history who might otherwise have trouble getting a loan."[207] In that regard, the ICO notes "It is not necessarily the case that there is discrimination because a person belongs to a particular social group, but they are being treated in a certain way based on factors that they share with members of that group."[208]

Further, discriminatory decisions can be based in factually incorrect or misleading information that can be the result of applying algorithms to large data sets that produce linkages where they might otherwise not be found, or alternatively create "spurious relationships". Spurious relationships refer to connections between data points that are the result of chance, not an underlying cause. The social repercussion of this form of digital discrimination can be exacerbated when data is recognised to create spurious relationships. These spurious relationships can become detrimental, particularly for consumers if decisions about the provision of goods or services, e.g., health or financial services, are based on such linkages.[209] This also raises important ethical considerations in relation to those practices. For example, big data companies must consider whether they can ethically use information when they cannot be certain that the results do not reflect incorrect linkages. This is particularly the case when this information relates to an identifiable person. The cart blanche application of algorithms to large data sets is also considered to be an ethically questionable practice as

[203] Ibid., p.20.
[204] Finn and Wadhwa, op. cit., 2014, p.7.
[205] Ibid., p.9.
[206] White, Martha, C., "Could that Facebook 'Like' Hurt Your Credit Score?", *Time, 14 June 2012. http://business.time.com/2012/06/14/could-that-facebook-like-hurt-your-credit-score/*
[207] "The 'Social' Credit Score: Separating the Data from the Noise", *Knowledge@wharton,* Wharton University, Pennsylvania, June 2013. http://knowledge.wharton.upenn.edu/article/the-social-credit-score-separating-the-data-from-the-noise/
[208] UK Information Commissioner's Office, op. cit., 2014, p.17.
[209] Fule and Roddick, op. cit., 2004.

"algorithms may be used in a way that perpetuates stereotyping or even bias."[210] The ICO suggests that practices, such as targeted or personalised advertising, that provide a tailored experience for users based on a customer's preferences, purchases and search history expressed online, they can also mean that they are being profiled in a way that perpetuates discrimination.[211] Relevantly, racially discriminatory practices were observed by Sweeney[212] who undertook an experiment during which she recorded the number of arrest suggestions that were linked with names predominantly given the black babies in America. Sweeney searched 2,000 "racially associated names" on google.com and reuters.com to determine whether names more commonly assigned to black babies turned up ad results that indicated criminal records. Although Google denied having any algorithm that incorporates racial profiling[213], that study concluded that so-called-black-identifying-names were significantly more likely to be accompanied by text suggesting that person had an arrest record, regardless of whether a criminal record existed or not. In addition, Rabess suggests: "Big Data is […] imperfect, because all of this data and the decisions we make with it are never completely objective."[214] Another example of an algorithm that can be said to reflect discriminatory attitudes is an algorithm used by St George Medical School to interpret medical data. That algorithm reflects discriminatory attitudes towards women and minorities as Rabess observed:

> The idea was to reduce variability and increase objectivity, but instead the school inadvertently institutionalized bias against women and minorities. This happened because it was relying on historical admissions data, which unduly favoured white male candidates.[215]

Thus, big data technologies and practices, including profiling and tracking, are susceptible to causing discrimination as a result of subjectively held biases that are inadvertently incorporated into the technological processes associated with big data. For example, Rabess suggests:

> When we translate cultural clichés and stereotypes into empirically verifiable datasets we introduce subjectivity into a discipline that strives for objectivity. When we imbue our Big Data insights with our race-based biases we project our prejudices onto subsequent observations. It's inevitable.[216]

Sweeney suggests that ads linking a person's name with criminal activity risk harming his or her reputation by suggesting wrongdoing when there is none.[217] This is also relevant to prospective employment applications where applicants become the subjects of online searches for information that impacts upon decisions to employ applicants.

In addition, big data practices that track, profile and collect vast amounts of information about individuals for commercial gain, and at times without the knowledge and consent of data subjects, raises the issue of exploitation which compromises ethical values. Exploitation is an

---

[210] UK Information Commissioner's Office, op. cit., 2014, p.16.
[211] Ibid.
[212] See Sweeney's report on her experiment at Sweeney, Latanya, "Discrimination in Online Ad Delivery", 28 January 2013. http://arxiv.org/ftp/arxiv/papers/1301/1301.6822.pdf
[213] Bosker, Bianca, "Google's Online Ad Results Guilty of Racial Profiling, According to New Study", *Huffington Post,* 2 May 2013. http://www.huffingtonpost.com/2013/02/05/online-racial-profiling_n_2622556.html?
[214] Rabess, Cecelia, "Can Big Data be Racist?", *The Bold Italic",* 31 March 2014. http://www.thebolditalic.com/articles/4502-can-big-data-be-racist
[215] Ibid.
[216] rabess, op. cit., 2014.
[217] cited in Bosker, op. cit., 2013.

ethical issue especially where individuals do not understand the extent of their participation in the big data industry or when they are unable to provide meaningful consent to these practices, and can ultimately have ramifications for society. For example, Finn and Wadhwa illuminate the practice of IT companies selling consumer data profiles to advertising agencies to enable them to more accurately target advertising campaigns.[218] Such practices highlight the unequal position between consumers and big data companies: "the surfing, browsing and purchasing 'work' in which consumers engage generate profit for powerful corporations, with no economic benefit for consumers themselves aside from access to free services."[219]

Exploitation can occur in any number of the technology processes used for big data. Once the data collector has the data, they can unethically, exploit that data. When users are exploited through data practices such as profiling, that exploitation also encompasses concerns over dignity, accuracy and discrimination.[220] We previously mentioned the use of big data processing techniques to identify credit worthiness and the fact that this practice potentially leads to discrimination. Additionally, exploitation is occurring is in relation to using information gleaned from big data sets relating to an individual's personal financial circumstances. This is a complex emerging practice:

> These start-ups hope to exploit a perceived shortcoming in traditional loan criteria based on FICO credit scores, in which people who have missed payments or lack borrowing experience would automatically be considered risky bets and penalized with higher interest rates on their loans. Or they could be turned down altogether, regardless of any mitigating circumstances such as a medical emergency or recent immigration to America. Lenddo, Neo Finance and Affirm make money principally through fees or commissions charged for each transaction. But whether or not their business models will last in the long run is another matter.[221]

Hong Kong-based Lenddo takes it one step further by using a debtor's social connections to exert pressure if he or she defaults on payments. For example, the start-up will tell customers' Facebook friends if they have not paid, and the friends' Lenddo scores could suffer if the customer fails to repay the loan.[222] Whether this practice has any long term viability is not yet certain but its disregard for ethical values of privacy in the exploitation of customers is clear: "It's the Wild West … like the early days of FICO."[223] Another example is provided by ZestFinance's uses of Google-like search algorithms to assess a person's credit risk by scouring thousands of potential credit variables.[224] Whilst this has benefited the organisations employing these practices, ethical issues arise in light of the potential for decisions to be made along subjective lines.

However, although big data practices such as tracking and profiling produce negative externalities such as discrimination, exploitation and/ or manipulation, such big data practices can also produce positive externalities.

---

[218] Finn and Wadhwa, op. cit., 2014, p.10.
[219] Ibid., p.11.
[220] Zimmer, Michael, "AOL Search Log Profiles Unmasked", *MichaelZimmer online,* 9 August 2006. http://www.michaelzimmer.org/2006/08/09/aol-search-log-profiles-unmasked/
[221] The 'Social' Credit Score: Separating the Data from the Noise", *Knowledge@wharton,* Wharton University, Pennsylvania, June 2013. http://knowledge.wharton.upenn.edu/article/the-social-credit-score-separating-the-data-from-the-noise/
[222] "The 'Social' Credit Score", op. cit., 2013.
[223] Cited in The 'Social' Credit Score", op. cit., 2013.
[224] Ibid.

Positive externalities of profiling and tracking

A number of positive externalities can be associated with tracking and profiling techniques have also been recognised. For example, data analysis can assist in identifying new trends. This can be beneficial in the health sector and ultimately for society as a whole. One of the most noted examples of now-casting is a service known as Google Flu Trends. By tracking the incidence of flu-related search terms, this Google spinoff service claimed that it could assist in identifying possible flu outbreaks one to two weeks earlier than official health reports. Early results released by Google suggested that when Google data are correlated with actual flu cases compiled by the Centres for disease Control, its estimates were 97 per cent to 98 per cent accurate.[225] However, the accuracy of that claim has been claim has been severely criticised.[226] Other trends such as unemployment or cultural tends can also be indicated by search engine use data, as well as a number of other data sources that are able to provide a snapshot of likely and emerging trends through the utilisation of data profiling:

> There are many types of real-time data streams that can now be assembled and analysed. Besides search engine queries, data for credit card purchases, the tracking and shipping of packages, and mobile telephone usage are all useful bodies of information.[227]

However, it remains relevant that the accuracy of results is not always certain. Nevertheless, the social benefits of such practices are observed by Manovich: "The rise of social media along with the progress in computational tools that can process massive amounts of data makes possible a fundamentally new approach for the study of human beings and society.[228] For example, "Today many more computer scientists are working with large social data sets; they call their new field 'social computing'."[229] According to the definition provided by the web site of the Third IEEE International Conference on Social Computing in 2011, social computing refers to "computational facilitation of social studies and human social dynamics as well as design and use of information and communication technologies that consider social context." ('Social Computing.')."[230] Thus, big data practices contribute to the rapid development of consumer and societal benefits, as well as having the potential to cause discrimination if big data practices are implemented without consideration for social and ethical values.

Therefore, information technology practice such as profiling and tracking can lead to a form of digital discrimination. Such discrimination requires effective minimisation to limit the socio-economic repercussions for those discriminated against. In addition, profiling and tracking using big data can also exploit and manipulate users when commercial gain is had at the expense of the social and ethical values of individuals. Despite these negative implications, big data practices such as profiling and tracking can also produce societal benefits, especially in terms of identifying health trends and risks. However, recognition of

---

[225] Park, Alice, "Is Google Any Help in Tracking an Epidemic?" *Time magazine*, May 6, 2009, at http://www.time.com/time/health/article/0,8599,1895811,00.html.

[226] Hodson, Hal, "Google Flu Trends gets it wrong three years running", *News Scientist,* 13 March 2014. http://www.newscientist.com/article/dn25217-google-flu-trends-gets-it-wrong-three-years-running.html#.VAa4PmSSy3c

[227] Bollier, David, op. cit.,2010, p.21.

[228] Manovich, Lev, "Trending: the Promises and the Challenges of Big Social Data", *manovich.net,* 28 April 2011, p.3. http://manovich.net/content/04-projects/065-trending-the-promises-and-the-challenges-of-big-social-data/64-article-2011.pdf

[229] Ibid., p. 4.

[230] Manovich, op. cit., 2011. p.4.

the social and ethical issues, especially in relation to user feelings of exploitation, can produce a benefit for all users. Once users become aware of the implications of big data technologies and practices, they can better adapt their online behaviour to preserve their social and ethical values. Awareness of the potential ramifications of employing ethically neutral technologies and practices can also be vital for big data companies operating into the future. A degree of attention must be turned to the potential for discrimination and its subsequent effects to ensure long-term credibility and subsequently, the commercial viability of companies and organisations seeking to capture the benefits of big data industry. In addition to this, profiling and tracking results in the implementation of personalisation techniques that also implicate social and ethical issues.

Personalisation techniques that follow profiling and tracking

Whereas the reasons for transparent practices as a means of gaining users' trust are entrenched in social philosophy, other methods such as personalised or targeted advertising are more commercially motivated and are the culmination of profiling and tracking practices. This too has the potential to affect levels of trust held by users. Personalisation techniques such as personalised and targeted advertising also raise issues of exploitation and manipulation.

Big data companies can build trust through providing users with personalised advertising and a more tailored online experience by predicting user preferences and conveniently providing relevant information. If users are receptive to personalised and targeted advertising then they may feel more trusting of the websites that "know them". Alternatively, users can feel exploited by this overt form of commercialisation, which also arguably encroaches ethical boundaries. An example of the relationship between personalisation techniques, including targeted advertising, and trust can be seen by social media networks that adopt personalisation techniques to build user trust, and in turn retain users as members on sites for the growth and development of their businesses:

> An example of trust is seen with personalisation techniques used by digital companies, particularly social media sites. The only way that a site such as Facebook or Twitter can continue to make money and grow is to personalise their offering so completely that those users trust it so much that they give up even more layers of personal information which can then be sold to industry.[231]

Thus, gaining a user's trust through the implementation of personalisation techniques, also raises the ethical issue of exploitation because, as seen in the previous example, trust is sought by big data companies not on the basis that it is a social value, but rather with the view to exploiting users for their information that can be sold to advertisers or used to create profiles for personalised advertising. This is the case with the process of personalised ads are bid upon in real time by advertisers who pay based upon the perceived value of a potential customer, "The new world of mass personalization requires relationships built upon mutual trust […] and soon big everything – are going to have to work hard to earn that trust from each and every one of us."[232]

In addition, personalisation techniques such as personalised and targeted advertising can manipulate consumer behaviours. Consumer manipulation can occur through customising

[231] Peters, Brad, "The Age of Big Data", *Forbes online,* December 2012. http://www.forbes.com/sites/bradpeters/2012/07/12/the-age-of-big-data/
[232] Ibid.

online advertising to individual users based on analysis of large data sets from social media profiles, search query histories, websites visited, and actions of visitors to websites. In relation to social media, manipulation can be an intentional objective when observing user behaviour. There have been a number of instances where social media networks and others have processed data to study human behaviours and trends and manipulate behaviours. Recently, it was revealed that Facebook Inc. undertook a social engineering study that involved manipulating what users saw on their news feeds, so that Facebook Inc. could assess whether there was a correlation between what users saw and what they subsequently posted to determine whether there is a correlation between what users are exposed to on Facebook and their moods.[233] Intentionally altering peoples' moods has potential consequences for the user, and raises ethical concerns related to the intentional manipulation of human emotions. The ethical risk associated with this were heightened by the fact that Facebook Inc. sort to manipulate emotions without any regard for the pre-existing emotional state of the study subjects (who were participating without knowledge and consent). Bollier[234] notes the risks associated with such practices stating that big data "may give certain players, especially large corporations, new abilities to manipulate consumers or compete unfairly in the marketplace", which may impact privacy, civil liberties and other freedoms.

Further, Finn and Wadhwa identify manipulation and objectification as one of the ethical issue arising in relation to personalised advertising[235], the result of big data technological processes such as data mining. They observed that, "Customers […] may also become increasingly manipulated by personalised advertising approaches that seek to generate interest in a product and convert interest into purchases."[236] Further, with respect to personalised advertising as a form of manipulation, Further, Taipale remarks:

> A business with economic motives is driving the process of data-driven personalization, but consumers have far less knowledge of what is going on, and have far less ability to respond. The benefits of personalization tend to accrue to businesses but the harms are inflicted on dispersed and unorganized individuals.[237]

However, Marc Rotenberg, Executive Director of the Electronic Privacy Information Center, admits that there are two sides to personalisation, which provides both positive and negative externalities for consumers:

> When Amazon and iTunes use their databases of consumer purchases to make recommendations to prospective customers, most people welcome the advice. It may help them identify just the book or music that they want. On the other hand, "people start getting very uneasy when buying suggestions are made based on how much we know about this particular person, a practice that takes us into the realm of behavioral targeting" – the "myTiVO thinks I'm gay" phenomenon.[238]

Thus, features like personalisation allow rapid access to more relevant information despite presenting difficult ethical questions because they fragment the public in troubling ways.[239] This also means that users who tend to have their own inbuilt social and ethical codes lose

[233] Meyer, Michelle, N., "Everything You Need to Know About Facebook's Controversial Emotion Experiment", *Wired.com,* 6 March 2014. http://www.wired.com/2014/06/everything-you-need-to-know-about-facebooks-manipulative-experiment/
[234] Bollier, 2010, p. 2.
[235] Privacy is recognised as a key concern in relation to personalised advertising
[236] Finn and Wadhwa, op. cit., 2014, p.12.
[237] Cited in Bollier, op. cit., 2010, p.23.
[238] Ibid.
[239] Cited in Boyd and Crawford, op. cit., 2012.

their ability to apply their codes when big data technologies and practices have the ability to manipulate their behaviours and market participation without their knowledge. However, this does not need to be the case for big data companies acting in the future. Ethical and social values that mirror user codes can be incorporated into big data policies. In doing so, big data companies increase their viability through supporting an ethically sound and socially aware big data industry.

Therefore, personalisation techniques raise issues of trust and manipulation, Irrespective of whether big data collectors seek to build user trust through transparent practices or by employing personalisation techniques as discussed here, trust is a vital ingredient for a prosperous big data industry. Trust remains a central social issue of big data practices that also has ethical underpinnings. Ultimately, building trust remains important to the future of big data because without consumer trust, amounts of data provided by individuals may decrease. This can limit the potential for positive externalities to flow from big data if data is limited and in turn, hinder big data companies to capture benefits of big data, such as increase in transparency in scientific, commercial and government decision-making. This is especially true, as citizens across Europe do not necessarily trust public or private sector organisations to safeguard their data.[240] Further, manipulation of consumers is another issues that arises in relation to personalisation techniques, such as personalised and targeted advertising that can lead users to modify or change their online behaviours without realising that they are being manipulated to do so. This form of manipulation compromises the ethical value of respect. Practices that compromise social and ethical values require recognition so that more socially and ethically sensitive practices can be a focus of innovation of the future big data industry.

4.3.3 Re-use and unintended secondary use or sharing

Reuse and unintended secondary uses of big data can have social consequences and also raise ethical questions. The risk of this occurring is increased when those using or re-using large data sets cannot be certain of the data quality or accuracy, such as when incorrect linkages and spurious relationships are created[241]. This can be a result of the methodology used in collecting and analysing data sets, particularly when data is being re-used. However, negative social issues are often overlooked in favour of the recognised and potential benefits that flow from practices including data re-use, such as addressing social problems including trafficking and global poverty by mining data available through big data banks with data related to social issues.[242]

There are real potential benefits of reusing big data. Public sector agencies have made it clear that data are an important element of social innovation. The European Commission has legislated for the re-use of public sector information[243] and open data policies are actively encouraged within Europe for the purpose of making available data for re-use. Internationally, institutions such as the US government and the World Bank have made their data available to the public for mining and further use. Individuals are using the data to create innovations, mainly apps, to address a particular social problem.[244] Organizations have been created to

[240] See Deliverable 2.2 of this project.
[241] This was addressed above in relation to discrimination.
[242] Desouza, Kevin. C., and Kendra . Smith, "Big Data for Social Innovation", *Stanford Social innovation Review,* Summer 2014. http://www.ssireview.org/articles/entry/big_data_for_social_innovation
[243] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Re-use of Public Sector Information: Review of Directive 2003/98/EC.
[244] Desouza and Smith, op. cit., 2014.

help make better use of big data for social problems, such as DataKind, which matches scientists and statisticians with non-profits for pro bono data work to help overcome the shortage of technology personnel capable of handling big data projects. Globally, the world's actors are making efforts to use open data and big data to develop solutions to social problems in innovative and collaborative ways.[245] Further, the New Alliance for Food and Nutrition Security[246], which is aimed at fighting poverty, includes, in its ten-year plan, a number of technology and data-based initiatives. There are a number of benefits flowing from the big data industry and the practices and technologies used to extract those benefits.

However, negative social implications also arise in relation to data re-use that require recognition. For example, public agencies and a newspaper in New York used the freedom of information Act to obtain detailed personal information about gun owners and subsequently published that information in the wake of a mass shooting.[247] This re-use of information had a number potentially negative social implications such as assisting criminals in targeting vulnerable homes where firearms are not held, or enabling criminals to identify homes from which they could steal guns.[248] This also left gun owners open to attacks on their reputations through their inadvertent connection to the behaviour of a mass murderer. This is particularly concerning when data relied upon is not necessarily up to date or accurate. This highlights the importance of authenticating data, and ensuring its accuracy and reliability when that data are used for secondary purposes. For example, social media companies often release incomplete data sets, which can compromise the validity of research findings, leading to poor practices and results.[249] The Royal Society argues that big data may also create a "data gap", where data becomes divorced from the context in which it was collected or initially analysed.[250] This means that data must be contextualised within its associated meta-data, i.e., how the data were acquired, method of collection, how to use the data set, how the data have been selected and how they have been analysed.[251] These issues also have ethical underpinnings that relate to privacy, especially where data sets have not been adequately anonymised or where information is freely available via public organisations that have not perhaps considered the number of potential re-uses of the data made available. Such ethically compromising practices also produce social ramifications. For example, cultural data, such as texts, artifacts and objects also raise unique externalities related to the making publicly available of sensitive religious or cultural data, and encouraging the re-usage of this data by different communities.

On the other hand, unintended secondary use is, as it name suggests, an unintended re-use of data, and is increasingly becoming a social issue of concern. The negative implications of unintended secondary use of information is illuminated by the examples addressed above in relation to social data that is being used to determine creditworthiness or as a basis for the denial or credit applications. Unintended secondary use of data produces social ramifications for the data subjects but it can also compromise ethical values when information posted 'privately' by, for example, a social media user is re-used. This is the case when information is re-shared in the context of social media. For example, Twitter data sets contain over 2.7

[245] Ibid.
[246] Feed the future, "The New alliance for Food, Security and Nutrition", 2014. http://feedthefuture.gov/lp/new-alliance-food-security-and-nutrition

[247] Desouza and Smith, op. cit., 2014.
[248] Ibid.
[249] Boyd and Crawford, 2012, p. 668.
[250] Royal Society, 2012, p. 14.
[251] Royal Society, 2012, p. 14.

billion messages, 80 million user profiles, and a 2.6 billion edge social network.[252] Analysis has revealed a growing trend of unintended re-sharing/ use:

> Where users defeat Twitter's simple privacy mechanism of "protecting one's tweets" by simply retweeting a protected tweet. We have shown through quantative and qualitative analysis that these privacy-violating retweets are a growing problem. More than 4.42 million tweets exist in our corpus that exposes protected information. As twitter gains popularity over time, we see an increasing trend in the number of privacy-violating retweets. Although there are many users who are unaware that their private tweets are being broadcast to the public, there are some who are aware of this problem.[253]

Whilst, this obviously highlights the issue of possible privacy violations, it also highlights how a negative social externality is created when this privacy invasive practice subsequently erodes user trust. This can also lead to users modifying their online behaviour. When distrust arises as a result of the permitted violation and ineffectiveness and inaccuracy of such a privacy policy, users may not only be harmed by repercussion of the tweet they thought to be private, but they may also alter their online behaviour. This has the ability to distort the social interactions in the digital space and produce inaccurate or "sabotaged data". Further, this practice of re-sharing or re-use can cause greater social problems for users, particularly in instances where the re-use of re-share of their data impedes their ability to be considered for many socio-economic endeavours such as organisational membership or, as addressed above, lead to discrimination in employment relations. This is because social networks like Twitter are:

> Indexed by Google. Users will need to understand that their tweets can be cached and searched, future and current employers may be reading their tweets, and they have no way of monitoring exactly who can read their tweets.[254] In relation to the Twitter issue: Unfortunately, retweeting allows protected tweets to be made public without the approval of the original tweet's author. Our results illustrate that not only are one's tweets never guaranteed to be "protected," but for those who think their tweets are protected, they may have more information leaking into the public realm than they were aware of.[255]

It follows that unintended secondary use or sharing additionally raises the ethical consideration of consent in relation to truth, user control and data company power in big data practices. Boyd & Marwick use the example of researchers with access to information and suggest:

> Researchers have the tools and the access, while social media users as a whole do not. Their data was created in highly context-sensitive spaces, and it is entirely possible that some users would not give permission for their data to be used elsewhere. Many are not aware of the multiplicity of agents and algorithms currently gathering and storing their data for future use. Researchers are rarely in a user's imagined audience. Users are not necessarily aware of all the multiple uses, profits and other gains that come from information they have posted. Data may be public (or semi-public) but this

[252] Meeder, Brendan, Jennifer Tam, Patrick Gage Kelley and Lorrie faith Cranor, "RT@IWantPrivacy: Widespread Violation of Privacy Settings in the Twitter Social Network", *Paper presented at Web 2.0 Security and Privacy*, W2SP 2011, Oakland, CA, 2010, p.1. http://www.cs.cmu.edu/~bmeeder/papers/Meeder-SNSP2010.pdf
[253] Ibid.
[254] Meeder, Tam, Kelley and Cranor,, op. cit., 2011, p.1.
[255] Ibid.

> does not simplistically equate with full permission being given for all uses. Big Data researchers rarely acknowledge that there is a considerable difference between being in public (i.e., sitting in a park) and being public (i.e., actively courting attention).[256]

The re-use of data, whether it is intended re-use as supported by policies such as those mandating open access, or conversely, re-used as a result of unintended secondary use or sharing represents a social issue with ethical considerations that has the potential to produce a number of positive and negative externalities. Data re-use can produce greater scientific, commercial and government transparency but it can also result in outcomes based on incorrect information or cause people to be identified when they would otherwise wish not to be. Unintended secondary use or sharing is, by its nature, a compromise of ethical values such as privacy and consent, with social ramifications such as discrimination during employment application procedures and other spheres of users' lives.

4.3.4 Data access

Big data technologies and practices that are either not universally acceptable or that enable or restrict access to large data sets raises social issues relating to potential inequality of access to data. This can create a digital hierarchy where only a limited number of big data actors have the potential and means to access big data sets, and extract benefits of that access. Conover[257] observes:

> The current ecosystem around Big Data creates a new kind of digital divide: the Big Data rich and the Big Data poor. Some company researchers have even gone so far as to suggest that academics shouldn't bother studying social media data sets - Jimmy Lin, a professor on industrial sabbatical at Twitter argued that academics should not engage in research that industry "can do better".[258]

Further, Manovich notes, "Only social media companies have access to really large social data - especially transactional data. An anthropologist working for Facebook or a sociologist working for Google will have access to data that the rest of the scholarly community will not."[259] This translates into some companies restricting access to their data entirely, whilst others sell the privilege of access for a fee and others offer small data sets to university-based researchers. This can produce considerable inequality in the system: "those with money – or those inside the company can produce a different type of research than those outside. Those without access can neither reproduce nor evaluate the methodological claims of those who have privileged access."[260] However, at this juncture, it is relevant to consider whether this inequality is also borne out a system that promotes companies and organisations with expertise to handle data over others. Given the technical nature of certain data technologies and related practices that can produces errors in information and, or spurious relationships, there can be times when data access may require access only by companies and organisations with relevant expertise. However, aside from situations where acute skills and technical expertise are required to minimise potential misuse or distortion of data, data access can be still be detrimental for society when this inequality is so entrenched that it denies other opportunities to innovate or participate in the big data industry. For example, inequality of

[256] Cited in Boyd and Crawford, op. cit., 2012, p.21.
[257] Conover, M., "Jimmy Lin", *Complexity and Social Networks Blog*, July 2011. http://www.iq.harvard.edu/blog/netgov/2011/07/the_international_conference_o.html
[258] Boyd and Crawford, op. cit., 2012, p.23.
[259] Manovich, op. cit., 2011, p.5.
[260] Boyd and Crawford, op. cit., 2012.

access to data entrenched in the university system that supports only certain researchers over others can result in restricted findings in areas of research and development:

> It is also important to recognize that the class of the Big Data rich is reinforced through the university system: top-tier, well-resourced universities will be able to buy access to data, and students from the top universities are the ones most likely to be invited to work within large social media companies. Those from the periphery are less likely to get those invitations and develop their skills. The result is that the divisions between scholars will widen significantly.[261]

Further, inequality of access between genders raises a social issue of concern because it can reinforce the issue of gender related access:

> Significantly, this is also a gendered division. Most researchers who have computational skills at the present moment are male and, as feminist historians and philosophers of science have demonstrated, who is asking the questions determines which questions are asked.[262]

Thus, the difficulty and expense of gaining access to big data produces a restricted culture of research findings. Further, data access is also limited those with expert computational knowledge.

> At 200 terabytes – the equivalent of 16 million file cabinets filled with text, or more than 30,000 standard DVDs the current 1000 Genomes Project data set is a prime example of big data, where data sets become so massive that few researchers have the computing power to make best use of them. AWS is storing the 1000 Genomes Project as a publically available data set for free and researchers only will pay for the computing services that they use.[263]

Although, this again raises the question of whether there are circumstances that warrant inequality of access due to the nature, complexity or size of the data sets, the aforementioned example also highlights when inequality of access issues can be perpetuated when large data companies have no responsibility to make their data available, and they have total control over who gets to see it. This can also lead to the modification of research objectives. Big data researchers with access to proprietary data sets are less likely to choose questions that are contentious to a social media company if they think it may result in their access being cut. The chilling effects on the kinds of research questions that can be asked - in public or private - are something we all need to consider when assessing the future of Big Data.[264]

Inequality of access to data and the associated problems such as reinforcing gender access issues or the modification to research objectives that are based in fear of losing access to large data sets require recognition. Recognition is important because they limit the potential uses of big data that could otherwise have resulted in benefits for big data industry and for society. The overall effect is that whenever inequalities are explicitly written into the system, they produce class-based structures. Manovich writes of three classes of people in the realm of Big Data:

> Those who create data (both consciously and by leaving digital footprints), those who have the means to collect it, and those who have expertise to analyze it'. We know that

[261] Ibid, p.5.
[262] Cited in Boyd and Crawford, op. cit., 2012, p 23.
[263] Office of Science and Technology Policy, "Obama Administration unveils 'Big Data' Initiative: Announces $200 million in new R&D investments", Executive Office of the President, Washington DC, 29 March 2012.
[264] Boyd and Crawford, op. cit., 2012, p. 23.

> the last group is the smallest, and the most privileged: they are also the ones who get to determine the rules about how Big Data will be used, and who gets to participate. While institutional inequalities may be a forgone conclusion in academia, they should nevertheless be examined and questioned. They produce a bias in the data and the types of research that emerge.[265]

Therefore, limited access to big data sets limits the potential use of data driven changes for social progress. However, there may be some circumstances that warrant reduced access to data sets such as when the technical nature of the practices being implemented or the complexity and size of the data require expertise that is not held by all big data actors. Moving forward, this issue of access and the social hierarchy associated with big data requires address to enable a variety of actors to capture societal benefits and produce a variety of results that can lead to solutions for human and social problems.

4.3.5 Open data

The availability of large data sets to the public, either through open government data or commercial open data policies and initiatives can have privacy implications. Open access can be, by its very nature, nature privacy invasive. Privacy invasive practices implemented by big data companies is an issue that has garnered much discussion by scholars and ethicists, as well as forming the basis of European laws and international policies. Privacy is also recognised as a social peril of data technologies.[266] Privacy is thus an obvious ethical and social concern when considering open data practices, especially as it can produce negative externalities.

It seems that where there is big data, privacy risks arise, and the potential for threats to privacy are increased when data is made open in a manner that makes leads to the identification of users, either directly or indirectly. Privacy and ethics go hand in hand and although relevant laws address privacy and data protection, it remains an ethical issue for most. For example the AOL search log profile made it clear just how easy it is to identify someone despite the use of a pseudonym. In 2006, users were able to identify the authors of the searches published by AOL. One user was easy to identify: No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything." The identity of AOL user No. 4417749 became easier to discern the more search queries she conducted:

> It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.[267]

Another research project undertaken using Facebook profiles has shown again how easy it is to compromise privacy and de-anonymise information from large data sets:

> So, again, the lesson learned here is how disparate pieces of seemingly benign information can be pieced together to make an otherwise presumed anonymous piece of data identifiable. I did it here by quickly analyzing the codebook, reading a press

[265] Cited in Boyd and Crawford, op. cit., 2012, p. 24.
[266] Cited in Bollier, op. cit., 2010, p 23.
[267] Zimmer, Michael, "AOL Search Log Profiles Unmasked", *MichaelZimmer online,* 9 August 2006. http://www.michaelzimmer.org/2006/08/09/aol-search-log-profiles-unmasked/

> release, and watching a video presentation. The New York Times did it with the AOL search data release, and I'm sure someone will do it with this Facebook dataset.[268]

These examples are unsurprising. Boyd and Crawford make the following observation:

> Data involving human subjects will undoubtedly raise ethical questions and issues, especially in relation to the privacy of that data, and indeed the human providing the data. They make this observation when considering the release of social media data by Facebook Inc., where that data that was though to be anonymised, could easily be de-anonymised.[269]

Thus, an ethical dilemma arises in relation to privacy where it is uncertain whether people are appear to be comfortable parting with personal information when they do not fully understand how the information can be made open. This dilemma is present when users are required to trade their personal information for access to a service.[270] That scenario is also indicative of when privacy is entwined with the ethics of consent. Consent is another ethical value that can be compromised by big data technologies and practices. For there to be ethical consent, the consent must be meaningful, and the approach: "Give us your data or we won't serve you […]'' cannot be considered meaningful consent.[271] It has become common practice in that "Individuals are often faced with a denial of services as the only avenue through which they can withdraw consent."[272] Consent in the context of collection and use of personal data is addressed under Article 2 (h) of the EU Data Protection Directive 95/46/EC.

Another example of where open data can raises ethical issues such as privacy exists when big data technologies are used to access data for research purposes. In that scenario, "many ethics boards do not understand the process of mining and anonymising Big Data, let alone the errors that can cause data to become personally identifiable. Accountability requires rigorous thinking about the ramifications of Big Data, rather than assuming that ethics boards will necessarily do the work of ensuring people are protected.*[273]* This echoes the sentiment expressed by Boyd and Crawford: "Just because content is publicly accessible doesn't mean that it was meant to be consumed by just anyone".[274] Similarly, data mining generally, and not just in relation to open data for research purposes, has been criticised for compromising ethical values such as privacy. Knowledge discovery allows considerable insight into data. This brings with it the inherent risk that what is inferred may be private or ethically sensitive. The process of generating rules through a mining operation becomes an ethical issue when the results are used in decision making processes that effect people, or when mining customer data unwittingly compromises the privacy of those customers.[275] This is because of the sensitive nature of the information being mined and the subsequent privacy related infringements that can occur in the absence of any framework.[276]

Relevantly, one of the benefits of the wealth of literature that the implications that big data practices have on issues such as privacy is that it has assisted in raising user awareness of potential privacy violations. This has resulted in big data companies taking notice of the

[268] Ibid.
[269] See Boyd and Crawford, op. cit., 2012, p. 18.
[270] Zimmer, op. cit., 2012.
[271] cited in Wright, op. cit, 2011, p. 203.
[272] Finn and Wadhwa, op. cit., 2013, p.18.
[273] Boyd and Crawford, op. cit., 2012, p.19.
[274] Ibid.
[275] Fule and Roddick, op. cit., 2004, p.1.
[276] For a more detailed discussion of the ethics of data mining see: Ibid.

implications their practices have on users and incorporating ethical values such as privacy into practice policies. Recently, the ICO made the following observation:

> There is evidence that some companies are developing an approach to big data that looks to place it in a wider and essentially ethical context. As well as a possible competitive advantage in being seen as a responsible and trustworthy custodian of customer data, adopting an ethical approach will also go some way towards ensuring that the analytics complies with data protection principles.[277]

Big data technologies and practices that have implication for the privacy of user data have also gained attention from ethicists:

> ''Privacy... is now recognized by many computer ethicists as requiring more attention than it has previously received in moral theory. In part this is due to reconceptualisations of the private and public sphere brought about by the use of computer technology, which has resulted in inadequacies in existing moral theory about privacy."[278]

Therefore, big data practices such as opening data, and the mining of that data, highlight the potential compromise of ethical values such as privacy. Recognition of negative externalities that can flow from the implementation of technologies and practices that compromise ethical values such as privacy provide a warning for big data companies operating into the future. Thus, big data actors can harness the knowledge of a growing awareness of other social and ethical issues and consider how to implement practices and policies that take into account social and ethical values.

4.3.6 Other issues identified from literature

Big data solutions for ethical and socially aware big data practices

Ethical and social values are often compromised by big data information practices, such as those addressed by this report. However, the severity of the compromise or the negative externalities they produce can be minimised. The literature reveals a number of key steps towards minimising potential harm to users and the industry per se by recognizing social and ethical issues arising in relation to big data technologies and practices, such as the ones addressed in this report. Ethical issues are a factor in the emerging discussion regarding solutions, especially as certain issues, such as privacy, are considered by regulators and lawmakers seeking to promote ethical and lawful innovation and commerce. Increasingly, calls are being made for an ethical framework to sit alongside legal frameworks that address issues related to big data collecting, storage and analysis such as consent and transparency.[279] This is said to be necessary because:

> Legal frames are not always the best way to grapple with ethical concerns. Indeed, much of what's at stake isn't running afoul of laws per se, even if it is still making people uncomfortable. What is that individuals find unsettling about data analytics? At

[277] UK ICO, *Big Data and Data Protection*, ICO office, UK, 2014, p.4. http://ico.org.uk/news/latest_news/2014/~/media/documents/library/Data_Protection/Practical_application/big-data-and-data-protection.pdf

[278] See for example Brey, and Moor as cited in Wright, op. cit., 2011, p. 211.

[279] Data & Society Research Institute, "Event Summary: The Social, Cultural, & Ethical Dimensions of 'Big Data", *The Social, Cultural, and ethical Dimensions of 'Big Data' Conference,* New York, 17 march 2014, p.2. http://www.datasociety.net/pubs/2014-0317/BigDataConferenceSummary.pdf

> various times, participants noted the problems associated with data uncovering information we'd rather have hidden or making inaccurate assumptions about us.[280]

At a recent conference on, among other things, the social and ethical issues associated with big data, the attendees and speakers highlighted the difficulty with aligning ethical values with technology processes that are utilised to make sense of big data, and ultimately produce benefits. Two such barriers are the difficulty presented by finding ethical frameworks for handling data's predictive capabilities, its potential persistence, and its ability to move across various sectors and platforms, and second, how are ethical values translated into code because it is simply to difficult to incorporate value judgments into algorithms.[281] However, awareness for the need to consider ethical implications of big data practices and technologies is gaining ground. Davis and Patterson observe four common elements of a big data ethics framework as identity (the relationship between online and offline identity), privacy (who controls access to it), ownership (including can ownership rights be transferred and the obligations on users of data), and reputation. (how do we manage how we are perceived and judged). Thus, both individuals and organisations have a legitimate interest in how big data is handled. Thus, Solutions can be achieved through the implementation of transparent practices, more user-friendly privacy settings, and/ or the implementation of privacy by design. Further, risks can be detected and diminished by carrying out specific risk impact assessments, such as a privacy impact assessment. In addition, an ethical impact assessment can also be carried out in relation to data technologies as enunciated by David Wright. Wright suggests that an ethical impact assessment may be a solution to predict and combat the ethical risks associated with information technology, which is relevant to big data.[282] Furthermore, the carrying out of an ethical impact assessment to determine the ethical risks of particular information technology related policies, projects or programs is preferable to prescriptive rules because of the ambiguity of the notion of what can be deemed ethical or not when having regard to the context.[283]

Furthermore, recognition of social and ethical issues arising in relation to big data practices is occurring in practice. The ethical framework implemented by IBM and referred to earlier in this report is a useful example of how a big data company can capture the potential benefits of big data whilst recognising social and ethical issues. Whilst this framework promotes user privacy, it also aims to achieve transparency in its practices. In relation to the recognition of privacy provided by that framework, it is a step towards recognising the issue of privacy is as much an ethical issue as a legal issue. It also indicates that where the law fails, there is some hope that a common understanding of the ethics of big data will come into play to prevent abuses such as exploitation of data, particularly of personal data.[284] The ICO comments:

> It is significant that these frameworks have been developed not by regulators but by companies themselves, as a response to the situation in which they find themselves in the big data world. They are approaches that are derived from the relationship between a company and its customers, and consider how to put that relationship on an equitable footing, rather than being derived from the company's need to comply with statutory

[280] Ibid., p.5.
[281] Data and Society Research institute, op. cit., 2014, p.2.
[282] Wright, op. cit., 2011, pp.199-226.
[283] Ibid.
[284] See for example, Chessel, Mandy, "Ethics for Big Data Analytics", *IBM Big Data Hub,* 2014. http://www.ibmbigdatahub.com/sites/default/files/whitepapers_reports_file/TCG%20Study%20Report%20-%20Ethics%20for%20BD%26A.pdf

obligations.[285]

Therefore, possible solutions to the big data technologies and practices that impeded social and ethical values require recognition as a key ingredient to enabling the continued growth of the European big data industry whilst safeguarding the social and ethical values of society and its citizens.

4.4 CONCLUSION

Information technology practices associated with big data, including transparency, personalisation techniques, profiling, tracking, re-use, unintended secondary use, sharing, data access, open access and manipulation raise a number of social and ethical issues including trust, discrimination, privacy, and exploitation. These issues arise in relation to the aforementioned big data practices, and the technologies that support them, because they involve people, their characteristics and their behaviours. Furthermore, despite some of the negative implications of big data practices, they can also produce a number of positive externalities for individuals and society. However, as the big data industry is still novel, frameworks that preserve social and ethical values are not yet commonplace. This may also be a result of the fact that defining social and ethical codes for big data practices and technologies is difficult, especially when ethics and social values differ across cultures and peoples. Nevertheless, there is an increasing awareness that big data practices and technologies raise social and ethical issues. In addition, a number of social issues identified in this report have ethical underpinnings and vice versa.

There is an inherently social component to big data because information practices that deal with data provided by people who operate according their individual social and moral codes. Any compromise to social values by practices that erode user trust, such as a lack of transparency, and promote discrimination, such as profiling and tracking have the potential to automatically effect negative or positive social changes, as well as changes to an individual's personal circumstances. Further, big data practices such as those relating to re-use, unintended secondary use and sharing, access to data, open access and manipulation raise ethical issues such as privacy as well as exploitation. In fact privacy is an overarching issue for a number of big data practices. The practice of manipulation is also enabled by big data technologies, especially in relation to personalised or targeted advertising. Nevertheless, big data practices can assist social innovation and solve social problems.

Transparency can produce positive and negative implications for big data companies and users. On one hand, transparency builds user trust and in turn, promotes the disclosure of more data by trusting data subjects, whilst on the other hand, it can cause users to modify or distort their behaviour and limit the amount of data they provide or perform data sabotage. Ultimately, building trust remains important to the future of big data because without consumer trust, amounts of data provided by individuals may decrease. This can limit the positive externalities of big data for big data companies and organisations and in turn, hinder big data companies to capture benefits.

[285] UK Information Commissioner's Office, *Big Data and Data Protection,* ICO, UK, 2014, p.45. http://ico.org.uk/news/latest_news/2014/~/media/documents/library/Data_Protection/Practical_application/big-data-and-data-protection.pdf

Information technology practice such as profiling and tracking can lead to a form of digital discrimination. Such discrimination requires effective minimisation to limit the socio-economic repercussions for those discriminated against. In addition, profiling and tracking using big data can also exploit users when commercial gain is had at the expense of the social and ethical values of individuals. However, awareness of the negative externalities of big data practices such as exploitation can translate into positive outcomes for users who may motivate big data actors implementing policies that preserve ethical and social values. This is undeniable benefit for society that can also support a sustainable big data industry.

Reuse and unintended secondary uses of big data can have social consequences and also raise ethical questions. The risk of this occurring is increased when those using or re-using large data sets cannot be certain of the data quality or accuracy. The re-use of data, whether it is intended re-use as supported by policies such as those mandating open access, or conversely, re-used as a result of unintended secondary use or sharing represents a social issue with ethical considerations that has the potential to produce a number of positive and negative externalities. Data re-use can produce greater scientific, commercial and government transparency but it can also result in outcomes based on incorrect information or cause people to be identified when they would otherwise wish not to be. Unintended secondary use or sharing is, by its nature, a compromise of ethical values such as privacy and consent, with social ramifications such as discrimination during employment application procedures and other spheres of users' lives.

Big data technologies and practices that are either not universally accessible or that enable or restrict access to large data sets raises social issues relating to potential inequality of access to data. However, there may be some circumstances that warrant reduced access to data sets such as when the technical nature of the practices being implemented or the complexity and size of the data require expertise that is not held by all big data actors.

The availability of large data sets to the public, either through open government data or commercial open data policies and initiatives raises the issue of privacy. Open access can be, by its very nature, nature privacy invasive. Big data practices such as open data, and the mining of that data, highlight the potential compromise of ethical values such as privacy. Recognition of negative externalities that can flow from the implementation of technologies and practices that compromise ethical values such as privacy provide a warning for big data companies operating into the future.

The social and ethical issues raised by the aforementioned technologies underline the importance of recognising that big data technologies can be used in a socially and ethically responsible manner. The absence of any moral or social code in big data technologies and practices can impinge upon user morality. Once individual and corporation value systems are understood, they can be implemented in a manner that aligns big data technologies and practices with those value systems. Further, ethical and social values can assist big data companies to morally evaluate big data technologies and processes to create a big data industry that is sustainable and influential on society. For example, transparent big data practices can reflect the inherent social and ethical aspects of vast amounts of data about people, rather than continuing practices that mean data are collected, stored and analysed without data subject awareness or consent. This can minimise the potential for social and ethical issues arising in relation to opaque practices.

Looking to the future, big data practices are likely more sustainable where they recognise social and ethical values as they will enable the positive externalities of big data, such as improvements in health care and scientific developments, to be continually available for captured. This is because longevity of the big data industry could be limited if users modify their use and online behavior or simply restrict the amount of data they provide if they feel that data technologies and practices impede their social and ethical values.

Ultimately, big data practices can attempt to achieve the objective set by the big data actors implementing them, whilst, at the same time, safeguard ethical and social values. This in turn, can diminish the negative externalities relating to the social and ethical issues addressed in this report. The preservation of ethical and social values by big data companies and organisations implementing relevant practices and technologies may be one of the defining characteristics of a sustainable big data industry.

5 POLITICAL ISSUES IN BIG DATA

Stephane Grumbach and Aurelien Faravelon,
Inria
With contribution from Grunde Lovoll, DNV

5.1 INTRODUCTION

Big Data will increasingly impact politics at all levels of society, displayed on Figure 1 and in particular at

(i) The regional level, its organizations and administrations,
(ii) The national level, the governance and political institutions, as well as
(iii) The international level, and diplomatic relations between States.

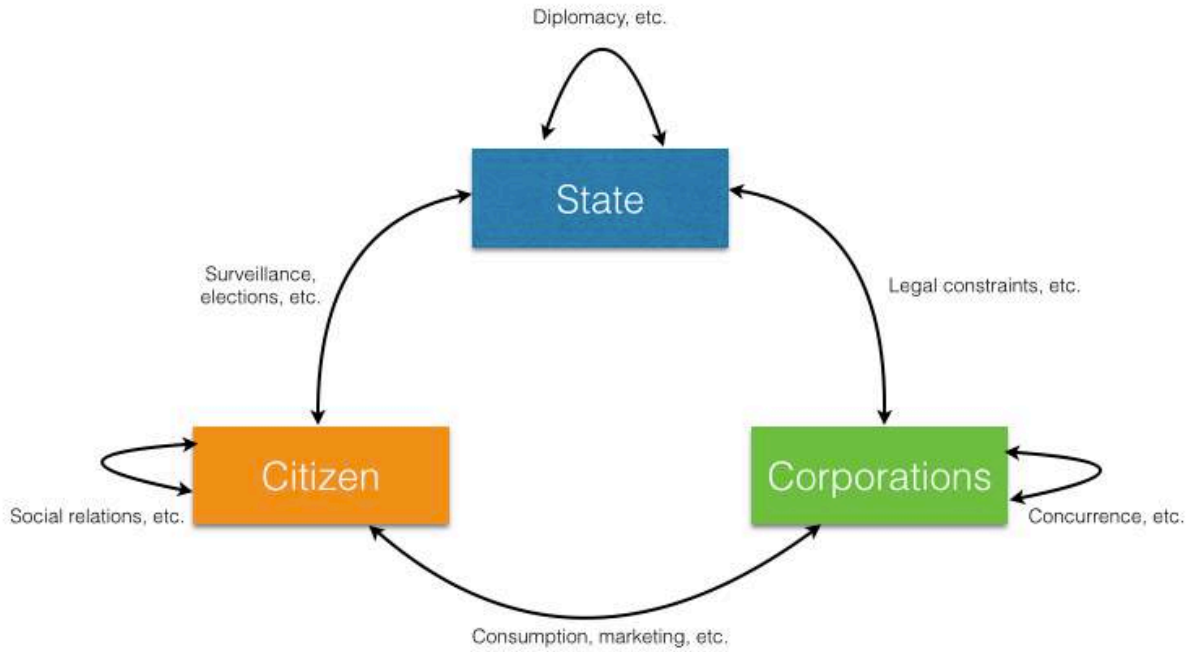
Big Data, and more generally the digital environment, contribute to change the balance and the relations between citizens, states and corporations. Our aim is to identify the main questions big data raise in each case. In this text, we analyze the externalities of big data on these relationships from a political viewpoint.

Corporations are currently challenging past equilibrium with states on many issues in a very broad spectrum of activity ranging from taxes to data protection, from utilities to copyright, etc. Corporations are challenging states, not only at home, but also more importantly remotely, in countries where they operate, in a way which generates new tensions, not seen with multinationals in other areas in the past.

Corporations challenge citizens. They offer them new services that become essential utilities. Most of these services are offered for free in a two-sided economic model. Corporations they keep the right to exploit the personal data of citizens, in a fast changing world, which has not established yet comprehensive rules of the game.

Citizens and states are facing each other in a fast evolving environment, where the available knowledge changes their relationship. On the one hand, data give states the ability to control citizens at scales that were unthinkable before. On the other hand, data allow citizens to demand more transparency from their governments.

The relations between individuals is booming on platforms where their life get materialized in the digital world supported by the services offered by the industry. Corporations compete to get direct access to their users with no middleman. Countries exchange data with at this stage little control, a situation that should evolve rapidly.



We show that data are the means to remotely control sectors of the economy and the society. Their impact is already visible, but the ongoing transformation of the society is only at its early stage. The changes imposed seem in many countries out of the reach of the political world, which often seem to have difficulties to cope with the challenges. The text is organized as follows: in section 2, we draw a brief overview of the political issues raised by Big data. The following sections go into further details. Section 3 inquire into the change in the relationships between citizens and states. Section 4 analyses the features and the status of digital services. Section 5 casts light on the economical effects of big data. Eventually, section 6 presents the geopolitical challenges of big data.

5.2 OVERVIEW OF THE POLITICAL ISSUES IN BIG DATA

Policy makers need to recognize the potential of harnessing big data to unleash the next wave of growth in their economies. They need to provide the institutional framework to allow companies to easily create value out of data while protecting the privacy of individuals and providing data security. They also have a significant role to play in helping to mitigate the shortage of talent through education and immigration policy and putting in place technology enablers including infrastructure such as communication networks, accelerating research in selected areas including advanced analytics and creating an intellectual property framework

that encourages innovation. Creative solutions to align incentives may also be necessary, including, for instance, requirements to share certain data to promote the public welfare.[286]

Critical questions exist - concerns over privacy in particular - about the implications of big data; who get access to what data, how data analysis is deployed and to what ends. In addition there are important questions about data access, discovery of "truth" from data, control and power. There are also important questions around various roles and classification and legislation around these: i.e. data creators, data collectors, data analysts and data keepers/stewards.[287] Businesses and governments are already exploiting big data, and in doing this they are often pressing the limits of legality, data quality, disparate data meanings and process quality. This can result in poor decisions, with individuals are bearing the greatest risk and the possible negative consequences; outcomes which could extend into social, economic and political realms (service denial, recovery cost in an asset liquidation, and false accusations of engagement in terrorism). The legality of the associated collection activity, the disclosure, the consolidation and the mining of the consolidated database might be resolved, asserted or merely assumed[288]. Also; both copyright and data ownership rights are difficult to enforce because of the difficulty of tracking the lineage of data streams once the resource is made available[289]. Hence, a big data governance programme needs to address the challenges associated with sharing of data between applications/actors and compliance with geographical trans-border data regulations, along with strong protection requirements for personal, health, and financial data[290]. In addition, such a programme also needs to take the type of data (i.e. Global Positioning System-GPS, credit history data, internal structured data such as Customer Relationship Management-CMR or inventory data, sensor data, external unstructured data such as Facebook or Twitter posts as well as internal unstructured data such as text documents) into account since provider/user obligations may differ for each category.[291]

It is noted by Kaisler et al. that the main concern of governmental agencies is the lack of tools and trained personnel to properly work with big data, as well as the new set of privacy incursions, invasive and unwanted marketing.[292] Other emerging challenges are "quantity vs. quality" and data ownership under which comes a modicum of responsibility for ensuring data accuracy, although such an assumption (much less a policy) is extremely difficult. Clearly, some big data (i.e. business intelligence, financial, personal, etc.) must be secured with respect to privacy and security laws and regulations, whereas five levels of increasing security have been suggested: privacy, compliance-driven, custodial, confidential and lockdown. Apparently, not all data is created equal; some data is more valuable than other

[286] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Hung Byers, A., "Big Data: The next frontier for innovation, competition, and productivity", McKinsey Global Institute, 2011.
[287] Boyd, D. and Crawford, K., "Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon", *Information, Communication & Society*, Vol.15, No.5, 2012, pp. 662-679.
[288] Wigam M.R. and Clarke, R., "Big Data's Big Unintended Consequences", *Computer*, Vol.46, No.6, 2013, pp.46-53.
[289] Koutroupis, P. and Leiponen, A., "Understanding the value of (big) data", presented at the 2013 IEEE International Conference on Big Data (IEEE BigData 2013), Santa Clara, CA, USA, October 6-9 2013, pp. 38-42.
[290] Malik, P., "Governing Big Data: Principles and practices", *IBM Journal of Research & Development*, Vol.57, No.3/4, 2013, pp. 1-13.
[291] Ebner, K., Bühnen, T. and Urbach, N., "Think Big with Big Data: Identifying Suitable Big Data Strategies in Corporate Environments", presented at the 47th Annual Hawaii International Conference on System Sciences (HICSS), Waikoloa, HI, USA, January 6-9 2014.
[292] Kaisler, S., Armour, F., Espinosa, J.A. and Money, W., "Big Data: Issues and Challenges Moving Forward", presented at the 46th Annual Hawaii International Conference on System Sciences (HICSS), Maui, HI, USA, January 7-10 2013.

data – temporally, spatially, contextually, etc. Some compliance challenges may be regarded as follows:

- What rules and regulations should exist regarding combining data from multiple sources into a single repository?
- Do compliance laws apply to the entire data warehouse or just to those parts containing relevant data?
- What rules and regulations should exist for prohibiting the collection and storage of sensitive data either centralised or distributed?
- Should an aggregation of data be secured at a higher level than its constituent elements, or the other way around?
- Given data security categorisation, what percentage should reside in each category? What mechanisms will allow data to move between categories?
- What are the necessary requirements for making private and sensitive data anonymous? The availability of contextual data makes data masking much more difficult, and it should imply stricter requirements.

The side effects of big data technology and possible abuse by those who have access to it may create legal issues. As the technology becomes powerful, more importance must be given on governance and many problems are occurring due to the fact that the legal system cannot effectively catchup with the rapidly changing technology. Controversies regarding governmental surveillance activities exposed political, social and economic issues for computer technology and became an opportunity for making the majority to recognise the need for understanding and controlling the use of computer technology. Thus, big data use in governmental activities and private corporate activities suggested solutions for the need to provide an environment in which computer technology could improve in harmony with requirements of the society; reduce side effects due to abuse of technology, and help find the way for computer technology contribute to the mankind. Since more information than ever is known about individuals including sensitive data such as health, financial and insurance records, it is demanded that privacy issues are respected. Although openness and sharing of big data can be a critical lever and enabler for improved performance, assurance has to be provided that the value of the consent to share data will by far outweigh the potential risks involved. Privacy is fundamental to building trust relationships with customers, business partners, employees, governmental agencies and other stakeholders, where communication plays an important role on what is known as well as shared, how and what can potentially be used.[293] Therefore, the establishment of a regulatory framework would be a sensitive task that must align protection of users/consumers with obligations of businesses and providers. An open issue is inconsistent regulations and practices in the area of privacy protection due to different approaches in various sectors and variety of national legislation, in spite of a common series of principles, international treaties and rights guaranteeing efficient and internationally unified protection of privacy and ownership. Some (quite radical) proposals could be mentioned as follows:[294]

- Instead of informing competent authorities about any activity related to data security, companies and organisations will have to assume full responsibility for data;

[293] Park, C. and Wang, T., "Big Data and NSA Surveillance – Survey of Technology and Legal Issues", presented at the IEEE International Symposium on Multimedia (ISM2013), Anaheim, CA, USA, December 9-11 2013.

[294] Lovrek, I., Lovri, T. and Lučić, D., "Regulatory Aspects of Cloud Computing", presented at the 20th International Conference on Software, Telecommunications and Computer Networks (SoftCOM), Split, Croatia, September 11-13 2012.

- Companies and organisations have to inform the competent national authority, without delay, about any breach of data security;
- International rules will be applicable to companies and organisations processing data in foreign countries if the company or organisation offers its services to domestic market;
- Data governance aspects will authorise independent national authorities for data protection to punish companies and organisations breaching the international data protection law;
- The authorities will implement basic principles and data protection laws when cooperating with the police and judiciary in relation to criminal proceedings. This rule will be applicable to domestic and cross border data transfer;

Two rights are of particular importance for users:

- Right to "data portability": users will be given easier access to their data and they will be able to transfer their data more easily from one service provider to another;
- Right to "forget data": users will be allowed to delete data if they decide that they no longer need it.

In this section, we have outlined the main political challenges that big data raise. We have pointed out that values and rights, such as privacy are at stake. We have also underlined that the nature of social relationships change. We now provide a detailed analysis of these points.

5.3 RELATIONSHIPS BETWEEN CITIZEN AND STATES

The relationship between citizens and states already is strongly impacted by big data and will continue to be so. Three aspects are extremely important. First, big data reshapes the public space as more and more datasets are "open" - i.e. publicly available. Secondly, big data open a wide range of new possibilities in terms of governmental action and especially in terms of surveillance. Thirdly, data influence the relationship between citizens.

5.3.1 State transparency

State data are numerous. However, it does not mean they are easy to access, be it for technical reasons or because data sets remain secret. Nonetheless, states and political action may be more and more "transparent". States have started to open their data, i.e. make them publicly available, most of the time on online platforms.[295] Even when only little has been done by the governments, the wealth of publicly accessible datasets – interviews, reports, etc. - allows to know a great deal about the activity of people and administrations. Most of the time, state-related data (be it open data or data from other sources) is regarded as a tool to empower citizens with the ability to process state data and develop new services on top of them. For instance, *nosdeputes.fr* in France and *votewatch.eu* in Europe use publicly accessible data in order to assess the assiduity of members of the French or European parliaments. They present themselves as ways for citizens to evaluate the work of their representatives. Assessing politicians also means fact-checking what politicians say. Blogs, such as "Les decodeurs."[296] use publicly available data to do so.

Thus, the digitalization of the states and of the political life challenges the traditional balance of power between politicians and states and citizens. Indeed, citizens may judge in real time and permanently what politicians do. Politicians are thus constantly accountable for their

[295] See, for instance, http://www.data.gov in the US or http://www.data.gov.uk in the UK or http://www.data.gouv.fr in France.
[296] Les decodeurs, 2014. http://*www.lemonde.fr/les-decodeurs/*

action. This new balance raises questions about the nature and the efficiency of politics. Can politics be fully transparent or is a degree of opacity necessary to lead fruitful political actions? For instance, can sensitive negotiations happen without at least a certain amount of secrecy?

5.3.2 State surveillance

Data collection and data processing alter surveillance from a quantitative and a qualitative viewpoints. The amount of individuals surveillance bears on and the amount of data it produces are gigantic. Surveillance applies to pretty much anyone and not only people who represent a danger.

Small sets of individuals used to be the object of surveillance and state attention used to be costly. Edward Snowden's revelation has brought since 2013 to the world public opinion the scale of this surveillance, and the importance of developing principles on the use, which can be made from personal data. *The Washington post* presents a list of the NAS surveillance reports disclosed by Snowden.[297] Surveillance turns out to be less costly and easier as it relies on the cooperation with companies who possess billions of users. For instance, surveillance has changed with social media since it has become possible to control a whole population. As a result, the industry thus becomes an important actor of state surveillance. Data-driven surveillance also changes the very nature and goal of surveillance. Data collection and analytics turns surveillance in a predictive activity. The Swedish tax agency, for instance, relies on SAS's softwares in order to automate tax auditing and detect potential frauds.[298]

Clearly, surveillance will continue to evolve deeply in the coming decades. It will be incompatible with interests from the different sectors of our society. State security and privacy may be two examples of conflicting interests. On the one hand, state security demands intelligence and surveillance. In contrast, privacy demands that certain datasets remain unavailable. As state surveillance oversteps the boundaries of states (states may remotely spy on foreign people and communications), it may create geopolitical tensions. It should be emphasized that Europe occupies a very weak position on that debate, since in the absence of a strong European industry, any regulatory attempt, might end up involving merely international trade dispute arguments than a defense of specific values that Europe would like to enforce.

5.4 CITIZENS AND COMMUNITIES

Three main points of influence of big data on the relationships between citizens have drawn our attention. Big data impact political movements, social relationships between citizens, as well as the relationship to oneself.

---

[297] Todd Lindeman and Ashkan Soltan, "Communication breakdow", Washington post, July 2014. http://apps.washingtonpost.com/g/page/world/communication-breakdown/1153/

[298] SAS, "Predicting and preventing tax fraud with analytics", 2014. http://www.sas.com/sv_se/customers/swedish-tax-agency.html

Big data influences political movements. In the Western world, Obama's campaign in 2012 is well known for its extensive use of technology.[299] Not only did Obama and his team communicate through social media, but big data analysis also helped to identify potential electors and contact them. In the Arab world, several authors analyze the "arab spring" as an example of the use of social media in order to build a collective political agenda. Social media helped to organise political actions and avoid Internet censorship. Social media also empower citizens with the appropriate skills with a new political weight.[300]

Big data influences social relationships between citizens. Citizens expose themselves in several ways. For instance, they play online games, experience social interactions and share their interests or work experience. By doing so, they generate massive amounts of data. The existence of such data provokes concerns about the possibility of a continuous and invisible surveillance called "underveillance". Citizens are always under the scrutiny of each other[301]. Such data can also be used to guide users in reaching people they may be interested in. Match making from big data analysis has applications in online dating for instance and job recruitment, two fruitful economic sectors.[302]

Eventually, big data influences the relationship to oneself. The "Quantified self" (QS) movement promotes self-knowledge and enhancement through self-recording and the analysis of the resulting data.[303] If the QS only gathers a limited amount of users, it finds an echo in the development of several applications, which rely on self-tracking to improve one's performances, health, etc. Nike's Nike+ running app allows one to analyse their runs and compare their performances.[304] It also changes each run in a competition with other users. Toggl automatically records the activity periods of workers in order to help them enhance their productivity[305]. Eventually, several applications are dedicated to health management. They record sleep habits, heart rate, etc. Devices, such as smartphones may thus turn into health management devices. Following this trend, Apple has recently announced that its new iOS 8 would integrate a health application, gathering a user's health data and applications.[306]

Everyone could thus become a data scientist or a data journalist by producing and analyzing their data. If companies present self-tracking and analysis as a means to promote one's autonomy, some researchers doubt it will actually do and denounce the limits of seeing oneself as only a compound of data[307].

5.5 NEW SERVICES AND ESSENTIAL UTILITIES

[299] Rutledge P., "How Obama Won the Social Media Battle in the 2012 Presidential Campaign", The National Psychologist, 2013. http://mprcenter.org/blog/2013/01/how-obama-won-the-social-media-battle-in-the-2012-presidential-campaign/

[300] Saada, J.. "'Printemps arabes' et révolution de l'information : Le poids des nouvelles technologies dans les relations internationales. Note de recherche ministère des Relations internationales, de la Francophonie et du Commerce extérieur du Québec", http://dandurand.uqam.ca, 2013. http://dandurand.uqam.ca/uploads/files/publications/etudes_raoul_dandurand/2013_Saada_MRIFCE.pdf

[301] Quessada D., "De la suveillance", Multitude, 2010. http://www.cairn.info/revue-multitudes-2010-1-page-54.htm

[302] Kang Z., Xi W., Mo Y., and Bo G., "User Recommendations in Reciprocal and Bipartite Social Networks-An Online Dating Case Study", IEEE Intelligent Systems (EXPERT), Vol. 29, No 2, 2014, pp. :27-35.

[303] Quantified Self labs, "About", 2014. http://quantifiedself.com/about/

[304] Nike, "Nike+ running app", 2014. https://secure-nikeplus.nike.com/plus/products/gps_app/

[305] Toggl, 2014. https://www.toggl.com/

[306] Apple, "Health", 2014. https://www.apple.com/ios/ios8/health/

[307] Rouvroy A., The end(s) of critique : data-behaviourism vs. due-process, in Mireille Hildebrandt, Ekatarina De Vries (eds), *Privacy, Due Process and the Computational Turn,* Routledge, London, UK, 2012, pp. 143-167.

Essential utilities are services that are used on a regular basis by people or corporations, and have become indispensable for their activities. Water and energy supply constitute basic utilities. Utilities are regulated in essentially all countries to ensure the continuity of service, as well as its neutrality – i.e. it must be the same everywhere for every user. Utilities can be provided by public or private enterprises, sometimes enjoying some sort of monopolistic positions. Telecommunication means have become new utilities fully deployed in the second half of the 20th century.

Recent surveys, such as the iYogi enquiry on the place of technology in our lives, show that technologies, and especially digital ones have turned into a central utility.[308] As we possess more and more connected devices, (in the US, for instance, people posses in average 7 devices), connection turns out to be considered as necessary to one's quality of life. Internet, together with several of the basic services of the Internet, such as search facility, storage and communication, has become *de facto* essential utilities in recent years. Private corporations provide them and ensure their continuity and their neutrality, but with little legal environment yet. New utilities massively impact traditional sectors of the economy, by allowing new means to develop them, and in particular connect users and providers of services.

5.5.1 Digital services as basic utilities and public services

Public services in the digital arena are built as platform, which (a) provide services directly to their users, and (b) allow the industry to built services on the platform, through their API. The platforms support new services developed either by emerging developers, or by well-established industries, which have no choice to maintain their customer but to be compatible with dominant media.

At first, platform such as Facebook may have looked as trendy gadgets for a limited amount of users, such as students. However, large platforms ambition to have a number of users in the order of magnitude of the global earth population, and become de facto world monopolies. In some cases they have already succeeded. In some areas, such as California, telephone is considered as a basic and public utility as it strongly influences one's quality of life[309]. However, as the Internet, and more widely, digital means of communication plays an important role in business activities and socialization, the same can probably be said of them. Moreover, these technologies replace traditional tools such as the phone or TV as they offer concurrent services. If so, providing these utilities could be regarded as a public service and some thinkers, such as Susan Crawford, professor of law at Yeshiva University advocate this solution.[310]

Public services theorists, such as Louis Rollands, provide guidelines to identify public services and the constraints they should meet.[311] To Rolland, public services must satisfy a need of the general interest. Public services must exhibit three properties: (1) continuity, (2) mutability (they must follow the changes of general needs and interest and (3) equality. As a

[308] iYogi, "63% people prefer to stay connected than warm", 2014. http://www.iyogiinsights.com/research/63-people-prefer-to-stay-connected-than-warm.html

[309] California Public Utilities Communication, "General Communications Information", 2014. http://www.cpuc.ca.gov/PUC/telco/generalinfo/

[310] Cardozo Law, "Susan Crawford", 2014. http://www.cardozo.yu.edu/directory/susan-crawford

[311] Gilles Gugliemi, "Une introduction au droit du service public", 2014. http://www.guglielmi.fr/IMG/pdf/INTRODSP.pdf

matter of fact, services such as Internet providing or information search, respect these properties. Several internet service providers (ISP) provide continuity recovery plans. The Internet relies on a conceptual framework that emphasizes equality. The "net neutrality", theorized by Tim Wu[312], for instance, states that data should be treated equally without distinction between contents or users for instance. The neutrality of the Internet has been widely debated, and enjoys a legal framework. Several countries, such as Canada, enforce this principle. It serves the industry that offers service online allowing a free deployment of services and the emergence of new actors. In the absence of national actors, it serves the foreign actors of course. More could be done on the neutrality of some of the datasets constituted by companies, which currently have exclusive right to use them.

However, considering some digital goods as basic utilities or digital services as public services raises many questions. It questions the status of the involved data and the principles that govern their access. Should they also be public and open? It also questions the role of the State towards these new utilities. Crawford, for instance, claims that States should compensate the lack of private initiatives.[313] For instance, in rural areas, they should provide broadband Internet access when ISPs do not. Eventually, it also questions the legislations applicable to these services and their legal consequences. In France, for instance, public services are a matter for administrative authority and administrative law. In common law countries, common carriers are subjected to specific restrictions.

5.5.2 Public services in the digital arena

The growing digitalization does not only create new services, it also reshapes traditional basic ones. Smart tools allow to retrieve data and analytics are more widely used to answer contemporary challenges. Population growth, for instance, puts a strong strain on basic utilities management. Providing water – a limited resource – to an ever growing population or protecting this population from natural catastrophes such as earthquake is a challenge. Monitoring risks and optimizing basic utilities delivery is thus necessary. Such activities rely on digital tools such as power meters or water pumps that produce data. In France, for instance, ERDF – which delivers power – uses Linky, an intelligent power meter[314]. Data analytics is a means to optimize utilities consumption and save them when population lacks them. IBM, for instance, uses data analytics in Bangalore in order to save water.[315]

Some utilities management may also rely on Open Data. Risk management, as in the case of earthquake, does.[316] However, as in power delivery, most utilities management rely on proprietary data used to bill customers or for marketing purposes. As such data may be of public interest, for example in order to drive policy decisions, the status of data used to manage basic utilities must be questioned. As companies provide more and more services in order to make cities smarter and greener for instance, and as they retrieve data out of these

[312] Wu T., "Network Neutrality, Broadband Discrimination", *Journal of Telecommunications and High Technology Law,* Vol. 2, No 2, 2003 pp. 141-180.
[313] Sam Gustin "Is boradband access a public utility?", Time, Jan 09 2013. http://business.time.com/2013/01/09/is-broadband-internet-access-a-public-utility/
[314] ERDF, "Linky profile specifications", 2014. http://www.erdf.fr/medias/Linky/ERDF-CPT-Linky-SPEC-PROFIL-CPL.pdf
[315] IBM, "Bangalore Water Taps IBM for Big Data Analytics", 21 Feb 2014. http://www-03.ibm.com/press/us/en/pressrelease/43255.wss
[316] Global Earthquake Model, 2014.

activities, should they keep such data or open them, for instance to business partners and public instances.[317]

As in the case of digital services, some traditional services turn themselves into platforms by opening their data. Several energy company, for instance, do so.[318] In Transportation, the French national company provides a platform of open data and an API.[319] A similar trend exists across the world. The TransportApi API for example offers access to the data of London's transportation service.[320]

5.6 REGULATION AND GUARANTEE FOR SOCIETY

As for other industries, regulations are under development for online activities. But not only is the regulator slower than the developer, but moreover, this industry raises new challenges for political powers. The platforms are indeed playing an increasing role, in a de-territorialized world, where the legal competence of the political power is more and more challenged. First, the platforms disrupt traditional sectors of activity. Taxation is thus challenged by platforms which organize exchanges outside the traditional monetary framework. In addition, the risks associated with big data, due to either leakage of data, or misuse, which can never be excluded, raises serious concerns for governments and citizens. Eventually, states struggle to regulate the activity of companies which offer worldwide services and which are headquartered in foreign countries.

5.6.1 Taxation

Fiscal optimization is the first phenomenon that leverages taxation. Fiscal optimization designates the operation to choose the location of a company's headquarters according to the tax policy of a country. In Europe, for instance, Google's headquarters lie in Ireland because of the country's low tax policy. Most of Google's European activities are declared in Ireland. As a result, several other countries, such as France, protest against this situation because it prevents them from taxing the company's activities.[321]

Intermediation platforms also challenge taxation. They substitute laymen to well-identified and taxable entities such as hotels and tax drivers. Collecting taxes out of flat renting or carpooling is a hot topic. Several strategies coexist. AirBnB, which allows users to rent their homes, has been the object of inquiry from tax administrations in various countries. In the US, Airbnb is requesting the taxpayer information from.[322] The company collects itself the US taxes. In contrast, such a system does not exist elsewhere. In order to collect taxes, countries such as France condemn users who make money out of Airbnb.

Eventually, the status of digital currencies, such as bitcoins, remains an issue. The use of such currencies happens out of the regular banking system, and most transactions are anonymous.

---

[317] Siemens, "Smart city", 2014. https://www.cee.siemens.com/web/at/en/csb/CVC/Your_Industry/smart-city/Pages/smart-city.aspx

[318] Oliver Balch, "Can open data power a smart city revolution?" The guardian. http://www.theguardian.com/sustainable-business/open-data-power-smart-city

[319] SNCF, "Navitia-io", 2014. http://data.sncf.com/tools/navitia-io

[320] TransportApi, 2014. http://transportapi.com/solutions/

[321] Sam Schechner, "Google hit with huge French tax bills", The wall street journal, 2014. http://online.wsj.com/news/articles/SB10001424052702304788404579523863174616536

[322] Airbnb, "American taxes", 2014. https://www.airbnb.com/help/topic/157

States thus try to find new regulatory ways in order to control their value and force people to declare their revenues.[323]

5.6.2 Risks

In traditional economic sectors, risks have been measured and accepted politically. The State has designed appropriates bodies to take risks into account. This is the case for instance for energy management, food safety, drug and health administrations, etc. Nothing comparable exists for big data. Norms should be developed, which are under discussions. Norms could deal with data storage and data processing.

However, developing such norms demand compromises about the definition of concepts such as privacy and the definition of common means to assess risks and prevent them. In terms of privacy protection, for instance, such a compromise does not exist between the EU and the US as the EU considers that the US does not provide an adequate standard for data protection. Transferring personal data from the EU to the US is thus forbidden excepts for specific datasets regarding, for instance, air traffic.[324]

5.6.3 Conflicts between states and companies in a de-terrortorialised world

Tensions arise between companies who are headquartered in a country while operating and several areas of the world. Current tensions between European data protection regulators and Google, Facebook and US cloud services are clear examples of this example. In May 2014, the EU has open the door to a "right to be forgotten" on the Internet.[325] It has recognized citizens the right to ask search engine to remove links in order to protect their privacy. However, the range of such a right is questionable. Google, for instance, now provides users with a means to ask for the removal of links that endanger their privacy.[326] However, the search engine only removes links when users query one of its European sites. When users query sites, such as google.com, which do not fall under European law, it does not remove them. Eventually, such rights as the rights to be forgotten, must be balanced with freedom of speech and censorship.

5.7 BIG DATA AND ECONOMICAL DISRUPTION

Data-centric services reshape traditional non-digital services. They challenge their business models and sometimes threaten to pull them down. Most of the times, intermediation platforms – i.e. digital platform that directly connect users – challenge the usual social roles and the distinction between producers and consumers. The rise of platforms such as Uber (private driver hiring) or Airbnb (flat renting), for instance, show that digital services challenge many aspects of our economy.

---

[323] Audrey Fournier "Comment la France veut réguler le bitcoin", Le monde, 11 July 2014. http://www.lemonde.fr/economie/article/2014/07/11/comment-la-france-veut-reguler-le-bitcoin_4455225_3234.html#xtor=AL-32280270

[324] Export.gov, 2014. http://www.export.gov/safeharbor/

[325] Curia, "Arrêt de la Cour (grande chambre) du 13 mai 2014.", 2014. http://curia.europa.eu/juris/liste.jsf?num=C-131/12

[326] Google Support, "Search removal request under data protection law in Europe", 2014. https://support.google.com/legal/contact/lr_eudpa?product=websearch

The conflict between traditional taxi hiring models and new data-driven services is a good example of such a situation. Usual taxis have to buy state-regulated licenses. Clients can hire them by calling a devoted platform. Taxi companies usually are state-based or city-based. Uber is a company founded in 2009, which proposes a mobile app to book a car with a driver. The client can track the cars on their smartphone until it arrives. Uber is a global company, which offers services in several countries and its drivers do not have taxi licenses. Unlike traditional taxis, Uber does not have fix rates. They depend on several parameters such as the period of the day. The rise of Uber has provoked several protests. Taxis have accused the company of illegal taxiing activities.[327] In France, for instance, the conflict between regular taxis and Uber is a conflict about data access and sharing. Regular taxis do not want to open their data to potential concurrent and want to forbid geo-location for unlicensed taxis.[328] Belgium has banned Uber from its territory, with not much success.

A similar trend of conflict between traditional actors and new digital ones appears in sectors such as car pooling (with websites such as blablacar.fr or karzoo.eu), and hotel business (with websites such as airbnb.com). In each case, the development of data-centric services entails conflicts with existing services and legal procedures. The development of such services raises also problems in terms of tax collection. This development calls for the development of new business models for the industry and new interaction with the state services.

5.8 CONCENTRATION OF THE NEW ECONOMY

Digital services demand the collection of large amounts of data, and thus large amounts of users. Main platforms possess billion of users and built up *de facto* monopolies. To do so, companies develop strategies to collect and retain as much data as possible and prevent competitors from doing so.

Several technical choices allow collecting data, even from remote sites. Offering an API for instance allows developers to build services on top of a platform. It also enables the platform to collect data. Single-sign on tools, which allow users to connect to a wide range of services with a single account issued by a source service, is another means to collect data from the use of concurrent services. For instance, connecting to a service with a Facebook account gives Facebook the opportunity to collect usage data. “Social tools” such as the “like box” which allow users to like the content of a web site is another tool to capture user's data out of Facebook's websites.

As a result, some companies, such as Yahoo, tend to limit the access to their services to third party applications. Indeed, Yahoo has recently announced it would no longer allow single sign on to Flickr expect with a Yahoo account.[329] The company aims at banning third-party login (i.e. logging in with any account but a Yahoo account) from all of its websites. As for Flickr, several users disapprove this change either because they do not Yahoo to capture their data or because they see the necessity to use several accounts to access several services as

[327] MG Siegler, “Uber CEO: I Think I've Got 20,000 Years Of Jail Time In Front Of Me”, Techcrunch, 2014. http://techcrunch.com/2011/05/25/uber-airbnb-jail-time/

[328] Helene Bezmekian, “Une loi dégainée en vitesse pour apaiser la grogne des taxis”, Le monde, 9 june 2014. http://www.lemonde.fr/economie/article/2014/06/09/une-loi-degainee-en-vitesse-pour-apaiser-la-grogne-des-taxis_4434633_3234.html?xtmc=uber&xtcr=3

[329] Kristin Burnham, “Yahoo dumps Google, Facebook logins for flick”, InformationWeek, 6 june 2014. http://www.informationweek.com/software/social/yahoo-dumps-google-facebook-logins-for-flickr-/d/d-id/1269491

cumbersome (https://accounts-flickr.yahoo.com/help/forum/en-us/72157641974750055/page6/).

Buying competitors or companies with data sets or an expertise in data processing is another strategy to gain data and data processing power. For instance, TripAdvisor has bought LaFourchette in 2014. LaFourchette is a website dedicated to restaurants booking in Europe.[330] In 2013, Google acquired several big-data and analytics related companies such as DNNResearch (speech, vision and language understanding) and Wawii (machine learning)[331]. Acquiring promising companies is a matter of competition. Wawii, for instance, aroused the interest of Apple and Google. As companies acquire their potential concurrents, we notice a concentration of big data activities in the hand of a small group of large companies. Most of them are American and other countries do not have equivalent companies. As a result, their own start-ups cannot find national buyers.[332]

5.9 GEOPOLITICAL CHALLENGES

Data – and especially personal data – play a central role in the undergoing social revolution. They are a fundamental element of economical and political strategies and choices. Their production and exploitation lead to a new cartography of power relations. Internet giants, such as Google or Facebook, emerge. Subsequently, countries and geographic zones, such as the US, gather a major part of data storage and processing capabilities when others, such as Europe, are deprived from data and such capabilities. Data appear central in new geopolitical challenges. In this section, we underline that data flow and processing entail a new balance in the relations between States.

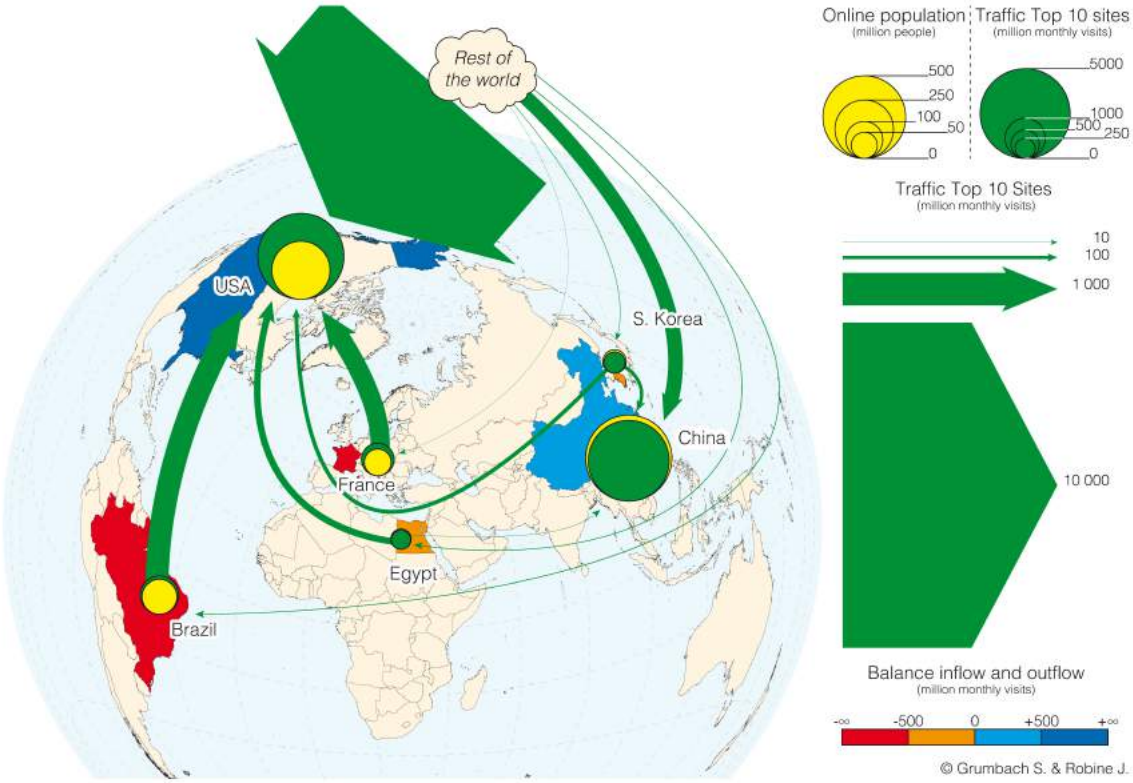
5.9.1 Locating data and data processing capabilities

Personal data production increases exponentially. The exploitation of these data in the last decade has led to the growth of the Internet giants, such as Google or Facebook, unprecedented in the history of industry. Yet, despite their importance in the new economy, neither online activity nor data flows are taken into account in traditional indicators, such as commercial balances, or raw materials price indexes. We use data flows metrics to measure this activity, which today stays essentially invisible. We have shown how the data collecting industry, based on the main intermediation platforms, works. These platforms occupy a dominant position in the world. In Europe for instance, they collect most of the personal data produced online, which is exported to the US. We have established a cartography of data flows, based on some representative countries, including France, Korea, Egypt, Brazil, and of course the US and China which are giants in the field. This cartography exhibits the knowledge asymmetries of the new economy.

[330] Challenges.fr, "Tripadvisor a finalisé l'achat de La fourchette", Challenges, 22 may 2014. http://www.challenges.fr/entreprise/20140522.CHA4159/trip-advisor-a-finalise-l-achat-de-la-fourchette.html

[331] Gregory Piatetsky, "2013 acquisitions in analytics and Big data", 21 august 2013, KDnuggets. http://www.kdnuggets.com/2013/08/recent-acquisitions-big-data.html

[332] Romain Dillet, "Dear Montebourg, Please Don't Kill French Startup Acquisitions" Techcrunch, 16 may 2014. http://techcrunch.com/2014/05/16/dear-montebourg-please-dont-kill-french-startup-acquisitions/



The global traffic is obtained from Trafficestimate, which is based on the number of monthly visits, and not the number of visitors, which generates subtle variations, but doesn't change the big picture. Traffic is measured in millions of visits. (The statistics of visits have been obtained the week of November, 4, 2013 from www.alexa.com and www.trafficestimate.com). Our cartography shows that the most visited websites are American and that the US collects most of the data produced in the Western world. It also shows the existence of regional actors, such as China, which gather large amounts of data. Eventually, it shows that European countries, such as France, exports most of its data to the US. The cartography clearly outlines the unbalance between the US and the EU in terms of data storage and the week position of the latter.[333]

5.9.2 Tensions and difficulties in data geopolitics

Europe is primarily characterized by marginal web sites that do not harvest large amounts of personal data, and many European countries are reliant on services provided in foreign countries such as the US, many of which are becoming as necessary for the economy as utilities such as transportation or energy. Such dependency generates tensions between regions that may result in political consequences.

[333] This cartography has been published in Frénot S., Grumbach S., "Les données sociales, objets de toutes les convoitises", *Herodote,* Vol. 1-2, No 152-153, 2014, pp. 43-66.

Laws and there spirit are a possible matter of conflicts. For instance, the EU directive on data protection – which all EU countries implement – emphasizes the protection of privacy. It forbids the transfer of data outside of EU in a country that the European Commission has deemed not to provide an adequate level of data protection. The US are not an approved country for data transfer outside of the EU.[334] In contrast, the US data discovery rules allow litigants to ask for the disclosure of data.[335] Forbidding such a disclosure may affect the run of US justice. The US emphasis on national security – and the subsequent disclosure of data it demands – is also in potential conflict with the EU emphasis on privacy.

Diplomacy is another field of potential tensions. Harvesting data is necessary for foreign policy making and intelligence for instance. The development of digital technologies has led to the intensification and transformation of tradition phenomena, such as spying. Some of them – such as PRISM, a US electronic surveillance initiative – are well documented. However, new activities, such as data visualization acquire a growing importance in foreign policy management.[336] The Internet accommodates large amounts of data that may be of interest in the field. For instance, such data may help to picture the political or legal situation of a country and inform the policy of other countries. Countries which centralize data visualization industrials thus possess a competitive advantage.

5.10 CONCLUSION

Political issues in big data revolve around the change of balances in relationships between states, corporations and citizens. Big data seriously impact the balance of power between politicians and citizens and between states and corporations. As intermediation platforms become central in economy and to build social relations, they become indispensable. As most of them are American, they challenge the geopolitical balance.

Indeed, the absence of corporations in the big data sector in Europe, and the increasing dependency upon US systems for services that can now be considered as utilities, restricts the capacity of Europe to react to legal disputes related to values Europe is committed to preserving. The recent conflicts on the right to be forgotten between US companies and EU show how hard it is to enact such values.

The data that are handled by US corporations fall under US laws, thus leading to some new territoriality of the American laws, despite the fact that the data originates from European citizens and may be stored on installations on the European soil. The US are not the only country to store their data on their own systems, China, Russia, as well as other (mostly Asian) countries have developed strong local systems, which harvest most of the data of their population. Europe is therefore in a position of weakness that exacerbates strong negative externalities of the big data industry, which are due to the very weak political capacity on its own territory. Moreover, there is a risk that such disputes may be framed as trade disputes rather than conflicts of law, and may be vulnerable to the exercise of economic and political

[334] European Commission, Commission decisions on the adequacy of the protection of personal data in third countries". http://ec.europa.eu/justice/data-protection/document/international-transfers/adequacy/index_en.htm
[335] Beirne, Maynard and Parsons, LLP, "Discovery vs. Privacy: When Laws Implementing The E.U. Directive on Data Protection Conflict with U.S. Discovery Rules", 12 january 2007. http://www.bmpllp.com/publications/78-discovery-vs-privacy-laws-implementing-eu-directive-data-protection
[336] Bronk C., Smith S., "Speaking out. How Data Visualization Can Change Diplomacy", *Foreign Service Journal,* 2012, pp.11-15. http://bakerinstitute.org/media/files/Research/77e62dd6/BRONK_030212.pdf

power through threats to cut vital services.

6 SUMMARY

Anna Donovan
Trilateral Research & Consulting

This report details a number of economic, legal, social and ethical and political issues that arise in relation to big data, particularly in relation to processing practices and technologies of big data. These issues are important because they illuminate areas where positive externalities may be captured, whilst also underlining negative externalities that require address.

6.1 ECONOMIC ISSUES

Big data implicates economic issues that have positive and negative social consequences. Big data relates to the economy because it can be a catalyst for innovation, in particularly when new business models require development to incorporate strategies for deriving the added value from big data and in order to capture the efficiencies of big data across a number of sectors. However, concerns for privacy are raised along side these positive effects.

Big data can boost the economy primarily because it increases efficiency and supports innovative business models, particularly in the following sectors: retail; manufacturing; health care; public; and life sciences. In terms of efficiency, big data can assist industry and other stakeholders in making "proactive knowledge-driven business decisions"[337] and the "extraction of embedded intelligence and data insights". Other business efficiency benefits include customer intelligence, supply chain management, quality management, risk management, performance management and fraud detection. Sector-specific efficiency benefits include healthcare efficiencies, reduced strain on infrastructure, better provision of energy, greater accuracy in prediction and measurement of weather events, as well as others. Improvements in efficiency might also support innovative business models by reducing entry barriers and making it less risky to launch new products, services or companies because of improved information and reduced uncertainty. Finally, consumers themselves may experience economic impacts through the provision of services at no cost, based on the value of the usage data generated by the service for the company. All of these discussions promise significant positive economic externalities in relation to big data. However, in terms of negative economic externalities, maintaining data subjects' privacy is one of the major obstacles for big data actors.

6.2 LEGAL ISSUES

Big data processing implicates intellectual property rights, namely copyright and sui generis database rights, the data protection framework, and raises concerns regarding liability, due process and jurisdictional issues.

Copyright laws also raise issues with contracting and licensing which are methods used to deviate from the standards imposed by the intellectual property rights framework. An example of when complexity or ill-fitting regulation arises when attempting to apply the current data protection legal framework to big data processing. Thus, a tension exists between

the data protection framework and big data processing, which requires address. However, in order to prevent this continuing, the principles of the framework require attention to better apply to big data processing. Moreover, in terms of privacy and data protection risks, some actors attempt to address privacy concerns by anonymising data, although the subsequent linking of data sets may result in the ability to re-identify individuals once disparate data sources are linked together. Big data processing also raises issues other issues concerning liability due process and jurisdictional issues. Due process is implicated because big data processing can be used to inform decisions about people or even as part of automated decision making. Possible application areas are in marketing and targeted advertising, insurance, credit lending and even security-related activities. This opens a wide area of problems, which are partially dealt with by the data protection framework but also raise issues covered by non-discrimination law, consumer protection, etc. Big data processing activities also involve issues of liability namely, who is responsible for which fault, and issues of jurisdiction, which laws apply and which courts can deal with the problem. All these aspects of accountability become more complicated with big data processing and cloud computing as underlying infrastructure. Again we notice that the legal frameworks were conceptualized in another technological environment and with other use cases as reference. In this section we consider liability, followed by a discussion on issues of jurisdiction.

Therefore, an examination of the legal issues that arise in relation to big data processing highlights the "gap" between technological capability and the legal framework, which means uncertain outcomes for economic competitiveness.

6.3 SOCIAL AND ETHICAL ISSUES

A number of social and ethical issues arise in relation to big data practices. Big data practices such as transparency, profiling and tracking, personalisation techniques, re-use, unintended secondary use, sharing, open data and open access implicate a number of social and ethical issues including discrimination, trust, privacy, inequality of access, exploitation and manipulation. This is because big data practices deal with data from people, and this human element reflects individual social and moral codes. These issues require recognition so that big data companies and organisations can incorporate fundamental social and ethical values into big data practices and policies. Ultimately, socially and ethically responsible big data practices can support the sustainability of a European big data industry. If big data practices compromise social and ethical values then data subjects may be reluctant to provide their data, or only to the extent that it gains them access to a service. This can limit the potential growth of big data.

Specifically, this report examines transparent practices that produce positive and negative implications for big data companies and users. On one hand, transparency builds user trust and in turn, promotes the disclosure of more data by trusting data subjects, whilst on the other hand, it can cause users to modify or distort their behaviour and limit the amount of data they provide or perform data sabotage. Ultimately, transparency is the key to building user trust, which in turn, leads to a greater amount of available data. Moreover, other information technology practices such as profiling and tracking can lead to a form of digital discrimination. Such discrimination requires effective minimisation to limit the socio-economic repercussions for those discriminated against. In addition, re-use, unintended secondary use and sharing of big data can also lead to social consequences and also raise ethical questions. The risk of this occurring is increased when those using or re-using large data sets cannot be certain of the data quality or accuracy. Big data technologies and practices

that are either not universally accessible or that enable or restrict access to large data sets raises social issues relating to potential inequality of access to data. However, there may be some circumstances that warrant reduced access to data sets such as when the technical nature of the practices being implemented or the complexity and size of the data require expertise that is not held by all big data actors. Finally, the availability of large data sets to the public, either through open government data or commercial open data policies and initiatives raises the issue of privacy. Open access can be, by its very nature, nature privacy invasive. Therefore, big data practices can compromise of ethical values such as privacy.

6.4 POLITICAL ISSUES

Lastly, a number of political issues arise in relation to big data. Big data will impact politics at all levels, namely: international relations between states; national governance and political institutions; and regional organisation and administration. An analysis of all levels is important because it identifies how the digital environment will change the balance between citizens, states and corporations. For example, corporations are currently challenging past equilibrium with states on many issues in a very broad spectrum of activity ranging from taxes to data protection, from utilities to copyright. This challenge, particularly when it occurs in remote countries, creates new tensions, not seen with multinationals in other areas in the past. The absence of corporations in the big data sector in Europe, and the increasing dependency upon US systems for services that can now be considered as utilities, restricts the capacity of Europe to react to legal disputes related to values Europe is committed to preserving. However, there is a risk that such disputes may be framed as trade disputes rather than conflicts of law, and may be vulnerable to the exercise of economic and political power through threats to cut vital services. Thus, BYTE will also integrate a geopolitical perspective to understand the political challenges facing the big data industry worldwide.

7 CONCLUSION

Anna Donovan
Trilateral Research & Consulting

The rise of big data processes and technologies raises a number of economic, legal, social and ethical, and political issues. The negative implications flowing from these issues require address to ensure that big data actors are in a strong position to capture the potential benefits of big data. Issues arise because of the nature of big data, that is because big data differs from traditional data because its amount is so large that it cannot be collected, stored, shared and analysed by traditional data analysis but requires new strategies and algorithms.

With respect to economic issues raised by big data, an assessment of the value propositions of big data, and an examination of big data and innovation, entrepreneurship and management efficiency support the notion that big data is to be regarded as a valuable asset. This means that changes to business models in a variety of sectors can be made to increase the value of the data as an asset whilst diminishing the negative externalities that arise such as invasions of privacy. Certain sectors that can benefit from changing business models to accommodate big data are the retail sector, manufacturing sector, health sector, public sector and the life sciences sector. Further, data markets and data warehouses will play an important role in the economics of big data.

Big data processing implicates intellectual property rights, the data protection legal framework and leads to problems with liability, due process and determining applicable jurisdictions. Copyright laws raise issues with contracting and licensing which are methods used to deviate from the standards imposed by the intellectual property rights framework. The protection by copyright and sui generis database rights of data sources clearly limits big data processing. It sets up isolated data sources, and making them available involves high transaction costs due to obtaining the necessary licenses for each source. Restricting the protection would give space for data flows and combination of data sets, as well as for new uses like data mining. Big data processing poses major problems for the data protection framework. Basic concepts, like the distinction between personal data and anonymous data, the data protection principles like purpose limitation and data minimization, and consent become difficult to sustain with big data processing. Also the rights of the data subject concerning transparency and access become more difficult to implement. Big data processing changes decision making and the construction of facts on which these decisions are based. This raises the question of how to sustain the autonomy and capability to act of persons. An important approach for this is looking how due process can be ensured in the context of big data processing. A key building block in such due process mechanisms is a larger transparency, which should also allow a larger insight in the logic behind the decision making. Big data processing relies on distributed computing and cloud computing is the enabling technology for that. Such cloud services can be combined and layered, leading to complicated architectures, which can be opaque. Also the technological convergence of services leads to problems in the application of legal frameworks, as these often were conceptualized on distinct use cases, which now get blurred. Both the opacity and the blurring of the application of legal frameworks creates difficulties for the application of liability

mechanisms, and makes the need arise for clarification or updating of these legal frameworks to the more complex technical environment. Similarly this creates jurisdictional problems, including a risk for the inflation of applicable laws, which raises the need for harmonization.

Social and ethical issues implicated by big data practices and technologies require recognition so that the negative externalities that flow from the implementation of technologies and practices that compromise ethical values such as privacy are diminished. A number of social and ethical issues arise in relation to big data practices. Big data practices such as transparency, profiling and tracking, personalisation techniques, re-use, unintended secondary use, sharing, open data and open access implicate a number of social and ethical issues including discrimination, trust, privacy, inequality of access, exploitation and manipulation. This is because big data practices deal with data from people, and this human element reflects individual social and moral codes. These issues require recognition so that big data companies and organisations can incorporate fundamental social and ethical values into big data practices and policies. Ultimately, socially and ethically responsible big data practices can support the sustainability of a European big data industry. If big data practices compromise social and ethical values then data subjects may be reluctant to provide their data, or only to the extent that it gains them access to a service. This can limit the potential growth of big data.

An examination of the political implications of big data reveals changing relationships between governments, citizens and corporations. An analysis of all levels is important because it identifies how the digital environment will change the balance between citizens, states and corporations. For example, corporations are currently challenging past equilibrium with states on many issues in a very broad spectrum of activity ranging from taxes to data protection, from utilities to copyright. This challenge, particularly when it occurs in remote countries, creates new tensions, not seen with multinationals in other areas in the past.

Therefore, a number of economic, legal, social and ethical and political issues are present in the big data landscape in Europe. These issues impact a number of sectors and potentially, the big data industry. Identifying these issues and understanding the positive and negative externalities they raise is key to the European big data industry moving forward. This is important because they have implications for both public and private sectors.

APPENDIX A – YOUTUBE LINKS ON ECONOMIC IMPACTS OF BIG DATA

1. The next chapter: big data and the future of the global economy
https://www.youtube.com/watch?v=9Ug0vtL8pjk
"Brad Peterson, chief information officer of NASDAQ OMX, Michael Flowers, analytics director for the Mayor's Office of Policy and Strategic Planning, NYC, and Bill Johnson, chief executive of Citi Retail Services, discuss how they are harnessing the power of data to change the realms of finance, banking and government to be smarter, more responsive and more efficient at The Economist's Information Summit on June 4th 2013 in San Francisco."

2. Royal Philips Bets on Big Data in Healthcare | Davos World Economic Forum
https://www.youtube.com/watch?v=6ts8g8Xp9Jo
"Royal Philips CEO Frans van Houten says its formed a new business group to bring data to your healthcare provider with the idea of helping improve diagnoses and delivery."

3. Science for the Green Economy Seminar Series - Big Data & Environmental Informatics
https://www.youtube.com/watch?v=ZsMd7dLRGx4
"This seminar, held on 14 May 2014, discussed how current research activities in environmental informatics are addressing this need, touching on approaches such as data mining, statistical interpretation, and predictive analytics for handling such 'big data'."

4. Economic Value of Big Data
https://www.youtube.com/watch?v=sgudQIEdKkA
"Author Paul Tallon expands on his article "Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost" and discusses how finding data governance practices that maintain a balance between value creation and risk exposure is the new organizational imperative for unlocking competitive advantage and maximizing value from the application of big data."

5. Education, healthcare & Big Data revolutionise AsiaPac economies
https://www.youtube.com/watch?v=L7o2Z25ZNUA
"Craig Mundie, Senior Advisor to the CEO, Microsoft speaks about revolutionary changes in the business environment in Asia Pacific economies: Education, healthcare, big data, access to technology are some areas he outlines as part of PwC's APEC 2013 CEO Survey."

6. Peter Sondergaard, Gartner, Says Big Data Creates Big Jobs
https://www.youtube.com/watch?v=mXLy3nkXQVM
"Peter Sondergaard, global head of research at Gartner, says by 2015, 4.4 million IT jobs globally will be created to support big data."

7. Creative Destruction, Technology and Big Data
https://www.youtube.com/watch?v=uOj75MJ77CM
"The rise of Big Data is bringing an enormous wave of economic growth as new businesses form. In this Daily Reckoning exclusive interview, we sit down with Harvard Business school's Clayton Christensen to discuss creative destruction in communication technology."

8. IBM Edge2013: Intel CIO Kim Stevenson on the "Sharing Economy & Big Data"
https://www.youtube.com/watch?v=VAHYixHu5nk
"Kim Stevenson, CIO of Intel discusses technology futures at the IBM Edge conference. Ms. Stevenson presented a session on how a modern enterprise operates in the "sharing economy."

She focused on how big data can drive competitive advantage and why people don't buy technology, but the benefits that it delivers."

9. Big Data to add £216 billion and 58,000 new jobs to the UK economy by 2017
https://www.youtube.com/watch?v=beRa85QOJ00
"Independent Media News discusses the impact of big data on UK economy in terms of employment and capital."

10. Big Data - Michael Chui of McKinsey Global Institute, at USI
https://www.youtube.com/watch?v=V3IBBUIYujs
"Michael Chui is a Senior Fellow of the McKinsey Global Institute. He is based in San Francisco, CA, where he directs research on the impact of information technologies, such as Big Data, Web 2.0 and the Internet of Things, on business and the economy. He co-authored the MGI report entitled "Big data: The next frontier for innovation, competition and productivity.""

11. 'Big Data'—The Digital Agenda for Europe and Challenges for 2012
https://www.youtube.com/watch?v=C_8UkMHUJcQ
"The EU's Digital Agenda champions the Internet as a means to achieve economic and social progress. In 2012, the Commission will be focusing on the potential of 'Big Data', the increasingly large and complex datasets that permeate the information economy. Mr Whelan's presentation outlined the Commission's plans in this area and put them in the context of other initiatives—on data protection, cloud computing, network security and the 'internet of things'—that the Commission is undertaking."

12. Leaders in Big Data
https://www.youtube.com/watch?v=8gMp0YC0_kM
"Discussing the evolution, current opportunities and future trends in big data Presented by Google and the Fung Institute at UC Berkeley."

13. Big Data & Analytics in the Retail Industry
https://www.youtube.com/watch?v=aZXVWOAcXjA
"This use case looks at how savvy retailers can use big data and analytics - combining data from web browsing patterns, social media, industry forecasts, existing customer records, etc. - to predict trends, prepare for demand, pinpoint customers, optimize pricing and promotions, and monitor real-time analytics and results."