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 D4.2:
Evaluating and addressing positive and negative societal externalities

Author(s):	Hans Lammerant, Paul De Hert, Vrije Universiteit Brussel Guillermo Vega Gorgojo, University of Oslo Erik Stensrud, DNV

Dissemination level:	Public
Deliverable type:	Final
Version:	1.0
Submission date:	31 December 2015

Table of Contents

EXECUTIVE SUMMARY

In this deliverable we evaluate the positive and negative societal externalities of big data. 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 4 (WP4), Evaluating and addressing positive and negative externalities. Previously the BYTE-project performed sectoral case studies making an inventory of positive and negative externalities in the use of big data, categorised as economic, social and ethical, legal and political externalities. These case studies were drawn from the domains of crisis informatics, culture, energy, environment, healthcare, maritime transportation and smart cities. The previous deliverable in WP4, D4.1, presented a horizontal analysis of the externalities uncovered in these case studies. This deliverable builds further on those results and presents an evaluation and analysis of best practices to address these externalities.

The concept of externality is originally an economic concept and using it outside an economic context poses some difficulties. Externalities are the effects on unrelated third parties of an economic activity. Such an externality is negative when this effect presents a cost for this third party, and positive when the effect is a benefit. Such financial translation in costs and benefits makes these effects comparable and an overall evaluation feasible. But when these effects cannot be translated straightforwardly in financial terms, comparison and evaluation becomes more difficult. In this deliverable we also consider effects of big data practices on third parties, which raise social and ethical, legal or political concerns. In these cases the problem at hand is not to make an evaluation of the overall effect in economic terms, but an evaluation of how different incommensurable objectives conflict with each other. Given these difficulties, the subject is approached through a widening use of the concept of externality. First we discuss economic externalities and best practices to deal with them. Here externality will be used in its general economic definition, as a positive or negative economic impact at third parties. In the final section of this part some non-economic externalities are also analysed from an economic perspective. In the following chapters the analysis will leave the confines of the economic definition of externality. Instead a new analytical frame to evaluate the non-economic externalities will be developed, based on how big data affects the interactions between actors. This analytical frame also makes visible how big data affects the conventional mechanisms or legal frameworks regulating these interactions. This helps explaining the appearance of the externalities perceived in the case studies. This perspective provides the elements for an evaluation of these externalities and of potential best practices in the further chapters.

Economic externalities

The economical externalities uncovered in the case studies are evaluated and corroborated with observations on the effects of big data outside of the case studies. The economical externalities treated are categorized into operational efficiency, innovation, new knowledge, business models and employment. Improved operational efficiency includes the capacity to make better decisions, to optimise resource consumption and to improve service quality and performance. Positive externalities in the innovation area include innovation through open data and enhancements in data-driven R&D through standardisation. The application of analytics and data mining techniques to large datasets can serve to gather

public insight and generate new knowledge. With the emergence of big data innovative business models can be found. However, also the lack of adequate big data business models, the challenge to traditional non-digital services and the potential dominance by larger players were raised as a negative externality. Lastly, an employment concern is the need for data scientists and analysts, while also the threat of employment losses for certain job categories that will be replaced by technology was raised.

After assessing the main economic externalities, the BYTE case studies and other big data initiatives were analysed in order to identify ways to diminish the negative effects and to amplify the positive impacts. As a result of this analysis, we propose a set of 8 best practices that are summarized in the following table.

Best practices	SMEs	Large corporation	Research institutions	Public agencies
Public investments in infrastructures				X
Funding programs for big data				X
Public investments in open government data				X
Persuade "big actors" to release some of their data		X		X
Promote big data in education policies				X
Look for interesting data sources	X	X	X	
Seek opportunities in new business models	X	X		
Create partnerships around data	X	X	X	X

The first two best practices are aimed to diminish the negative impact of scarce resources. Funding big data infrastructures is one way to bootstrap the adoption of big data, especially in such cases that require substantial investments. Public bodies can also dedicate resources to fund research and innovation programmes in big data. The following two best practices are intended to address the scarcity of data and are aimed to release new data sources that can thus be exploited. This concerns government data but also corporate data, although it remains difficult to develop proper incentives for corporate data sharing. The best practice about educational policies is aimed to reduce the negative impact of scarce data skills. Education policies have to address the current scarcity of data scientists and data analysts, but also promote the inclusion of data skills in a range of educational programs. Finally, the last three best practices are geared towards a change in the mind-set and boost the adoption of big data. Our case study research uncovered several opportunities for new business models, as well as opportunities for new uses of existing datasets. The creation of partnerships around data should be promoted to exploit the synergies of the different stakeholders.

Social and ethical externalities

Positive social benefits are similar to the positive economic impacts and can be captured with the same best practices as recommended to address economic externalities. Interoperability is a key factor in enabling big data practices. Investment in the several

dimensions of interoperability helps to make big data practices possible and capture the associated benefits.

An important negative externality is the risk for discrimination. Data mining can result in unintended discrimination due to choices in the problem definition and to discriminatory bias in the training data. Effective methods to deal with such bias have recently been developed and are object of further research. These methods can be integrated in auditing tools and in an anti-discrimination-by-design approach, similar as in privacy-by-design. The legal framework concerning anti-discrimination is less developed than data protection and mostly focusses on legal redress, but also establishes equality bodies working on mainstreaming of policies. They can play a role in addressing anti-discrimination in big data as part of these mainstreaming efforts. First by promoting the development of an anti-discrimination-by-design approach. Second by developing a transparency and accountability framework based on both the anti-discrimination legislation and the data protection framework. Coordination and cooperation between DPAs and equality bodies on this matter would be very useful. Lastly, big data creates new visibilities and makes it possible to discern between people on a whole range of health- and behaviour-related and other personal aspects. This also provides grounds for 'new discriminations'. The equality bodies will have to engage in a prospective societal and political debate on how to deal with these 'new discriminations'. This will have to lead to the development of new norms on which decision-making is acceptable based on these new visibilities.

Legal externalities

Several legal frameworks are not functioning well in the context of big data. The current frameworks do not scale in the context of a much higher amount of interactions, resulting from regular data flows used in big data practices, and lead to high transaction costs. Legal mechanisms requiring action for each individual transaction, like consent or authorisation, need to be substituted with aggregate or collective mechanisms. Further, the development has to be stimulated of 'by-design'-approaches, in which legal objectives are translated in technical requirements and taken into account during the design of technical systems.

Copyright and database protection: Evaluations of the current copyright and database protection laws show that this framework is too restrictive. Possible solutions involving legal change are to drop the sui generis-right on databases and to add new exceptions for data and text mining in copyright. A solution within the current legal framework is collective licensing, which lowers transaction costs but preserves remuneration. Extended collective licenses can make this framework more flexible. Still, the evaluations show that limiting the extent of copyright is preferable, as it leads to market expansion due to network effects.

Trade secrets: The case studies showed that companies can have 'privacy' problems as well, in the form of a need to protect confidential information. To address these concerns the adoption of a legal protection of trade secrets is recommended. Within such a framework stakeholders and standard bodies have to develop standardized solutions and a toolbox of legal, organisational and technical means which can be used to fine-tune data flows.

Privacy and data protection: The principles of the data protection framework can be implemented in a risk-based approach without limiting the extent of their protection. In such a risk-based application the actual risks associated with the overall set-up of technical, legal and organisational safeguards need to be evaluated, instead of the theoretical risks associated with anonymisation techniques or other technical safeguards. To make such overall evaluation possible, the development of standard solutions with known capabilities and risks for the whole range of technical, legal and organisational safeguards needs to be promoted. Further, privacy-by-design provides an important approach for mainstreaming privacy into the design process. It needs to be further developed and broadened from a purely technical approach to include also legal and organisation safeguards. However, it already provides a practical approach which can serve as an example to address other societal concerns. Lastly, individual transaction-based elements of the data protection framework, like consent, need to be substituted by more aggregated and collective mechanisms, which includes a strengthened role for DPAs.

Political externalities

The political externalities uncovered in the case studies related mostly to political economics. Potential problems concerning the relation of the private sector with the public and non-profit sector can be summarised as the fear for lock-in and dependency on dominant private players and the fear for rent-seeking by the private sector on public investment. The presence of positive network effects in the market for data and data services can easily lead to lock-in and captive value chains. Lock-in and dependency on dominant private players can be avoided by preventing the use of technical standards or platforms as gatekeeping and control mechanism. Tools are collective standard setting (instead of private), the use of open source, open protocols and open data, and data portability. However, the effect of open data can be ambiguous and strengthen the position of already dominant players. In this case capturing the monopoly rent through providing the data through licensing could be a better solution.

Restoring the regulating capacity of the state through unilateral protectionist measures or extraterritorial application of national legal frameworks can lead to legal conflicts or to barriers for big data practices. Although often difficult, international legal harmonisation is the best approach to address these problems. Big data can itself also be used to restore governmental control capabilities, but with risks for human rights infringements and political abuse. New legal safeguards need to be developed to restore the balance between citizens and state authorities.

This comprehensive evaluation of societal externalities of big data shows a broad agenda for policy-makers consisting of updating legal frameworks, the promotion of big data practices through public investments, and enabling policies and an active policy to keep markets open and competitive. Regulators and stakeholders also have an important role in developing tools and approaches to mainstream societal concerns into the design process and the big data practices. The recommendations summarises the agenda for policy-makers and other stakeholders, which can be derived from this analysis. In the appendix a methodology is presented for a quick scan of externalities and measures to address them at the level of an organisation.

1 INTRODUCTION

In this deliverable we evaluate the positive and negative societal externalities of big data. 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 4 (WP4), Evaluating and addressing positive and negative externalities. Previously the BYTE-project performed sectoral case studies making an inventory of positive and negative externalities in the use of big data, categorised as economic, social and ethical, legal and political externalities. These case studies were drawn from the domains of crisis informatics, culture, energy, environment, healthcare, maritime transportation and smart cities. The previous deliverable in WP4, D4.1, presented a horizontal analysis of the externalities uncovered in these case studies. This deliverable builds further on those results and presents an evaluation and analysis of best practices to address these externalities.

The concept of externality is originally an economic concept and using it outside an economic context poses some difficulties. Externalities are the effects on unrelated third parties of an economic activity. Such an externality is negative when this effect presents a cost for this third party, and positive when the effect is a benefit. Such financial translation in costs and benefits makes these effects comparable and an overall evaluation feasible. But when these effects cannot be translated straightforwardly in financial terms, comparison and evaluation becomes more difficult. In this deliverable we also consider effects of big data practices on third parties, which raise social and ethical, legal or political concerns. In these cases the problem at hand is not to make an evaluation of the overall effect in economic terms, but an evaluation of how different incommensurable objectives conflict with each other. This also makes differentiating between positive and negative effects difficult, as the negative effect for one actor will be the consequence of the positive effect for the other actor and vice versa. Depending on which interest is given priority, it will translate also in a negative or positive externality.

Given these difficulties, the subject is approached through a widening use of the concept of externality. First, in chapter 3 we discuss economic externalities and best practices to deal with them. Here externality will be used in its general economic definition, as a positive or negative economic impact at third parties. In the final section of this part some non-economic externalities are also analysed from an economic perspective. In the following chapters the analysis will leave the confines of the economic definition of externality. Instead a new analytical frame to evaluate the non-economic externalities will be developed, based on how big data affects the interactions between actors. This analytical frame also makes visible how big data affects the conventional mechanisms or legal frameworks regulating these interactions. This helps explaining the appearance of the externalities perceived in the case studies. This perspective provides the elements for an evaluation of these externalities and of potential best practices in the further chapters.

We start with a focus on economic externalities in chapter 3. In a first section the economic externalities uncovered in the case studies are evaluated and corroborated with observations on the effects of big data outside of the case studies. Based on this analysis a set of 8 best practices is proposed in order to capture the benefits of big data. They aim to diminish the negative impact of scarce resources, to address the scarcity of data and release

new data sources, to address the scarcity of skills and to change the mind-set in order to boost the adoption of big data.

Social and ethical externalities are addressed in chapter 4. The positive social impacts of big data are explained from the interaction perspective and linked with a couple of best practices. The negative social and ethical externalities concern equality and discrimination. An overview is given of how data mining can lead to discrimination and some methods to discover and present discrimination are presented. These methods can be integrated into an overall design anti-discrimination-by-design strategy, similar to and embedded in privacy-by-design. Further the role of equality bodies in driving the development of such design strategy and of an adequate accountability framework is put forward.

Chapter 5 deals with several legal externalities. It shows how the legal frameworks of copyright and database protection, of trade secrets and of data protection need to be updated to function effectively in the context of big data. This includes legal changes, which vary from limiting the extent of the framework in the case of IPR, establishing a new legal framework for the protection of trade secrets to substituting individual transaction-based elements for aggregated solutions in data protection. For data protection and the protection of trade secrets the deliverable highlights the importance of translating the objectives of the legal framework into a toolbox of legal, organisational and technical safeguards and an integrated design approach.

Chapter 6 deals with the political externalities. First the issue of lock-in and dependence in an economic context is analysed and measures to prevent such lock-in and to keep competitive markets are presented. Second the issue of dependence and the erosion of the state capacity to regulate and control is discussed. Lastly the need for new legal safeguards to prevent political abuse of big data practices is highlighted.

In our recommendations we present the agenda for policy-makers and other stakeholders, which can be derived from this analysis. In the appendix we present a methodology for a quick scan of externalities and measures to address them at the level of an organisation.

2 EVALUATING AND ADDRESSING ECONOMIC EXTERNALITIES

2.1 REVIEW AND EVALUATION OF ECONOMIC EXTERNALITIES

In this section we review the costs and benefits as well as the positive and negative economic externalities that were identified in the BYTE case studies, providing assessment of how to amplify the positive ones and how to diminish the effects of the negative ones. In order to do this, we complement the case study research conducted in BYTE with the experience and lessons learned of other relevant big data initiatives that were thoroughly analysed in Deliverable D1.3. The externalities and economic impacts analysed here are summarized in Table 1. We have employed the same categories as in deliverable D4.1 for organizing the economic externalities, i.e. operational efficiency, new knowledge, innovation, changing business models and employment.

Table 1: Summary of the economic externalities found in the BYTE case studies

Externality code	Externality description	Case studies					
		Crisis inf.	Culture	Energy	Environment	Healthcare	Smart cities
OPERATIONAL EFFICIENCY							
E-PC-BM-2	Better services	X	X		X	X	X
E-PC-BM-3	More targeted services	X				X	X
E-PC-BM-4	Cost-effectiveness of services	X	X	X		X	X
E-OC-TEC-2	Optimization of utilities through data analytics						X
E-PC-BM-5	Need of skills and resources	X				X	
E-OO-BM-7	Inefficiencies associated to big data				X		
NEW KNOWLEDGE							
E-PC-TEC-1	Gather public insight	X			X		
E-PC-TEC-2	Accelerate scientific progress through big data		X				
INNOVATION							
E-PO-BM-1	New products and services based on open data		X	X	X		
E-PC-DAT-1	Foster innovation from open government data		X	X			X
E-OC-DAT-2	Innovation through open data (and open source)	X					X
E-OC-DAT-1	Enhancements in data-driven R&D through standardization				X		
E-OO-DAT-5	Reluctance to open data			X			X
CHANGING BUSINESS MODELS							
E-OO-BM-2 E-PO-BM-2	Innovative business models based on big data			X	X	X	
E-OO-BM-1	Community building and innovation through big data			X			
E-OO-BM-8	Not clear data-based business models	X	X	X		X	X
E-OO-BM-3	Challenge of traditional non-digital services				X		X
E-OC-BM-8	Privatization of essential utilities	X					

E-OC-BM-7	Reduced market competition				X		
E-OO-BM-5	Competitive disadvantage of newer businesses and SMEs						X
E-PO-DAT-1	Open data puts the private sector at a competitive advantage				X		
E-OO-DAT-1	Digital divide				X		
EMPLOYMENT							
E-OC-BM-3	Data-based jobs			X	X		X
E-OC-BM-5	Employment losses for certain job categories						X

2.1.1 Operational efficiency

Improved operational efficiency is one of the main reasons for using big data, as reflected by externalities E-PC-BM-2, E-PC-BM-3, E-PC-BM-4 and E-OC-TEC-2. Here, data is used to make better decisions, to optimize resource consumption, and to improve service quality and performance. It is what automated data processing has always provided, but with an enhanced set of capabilities. Concerning better services (E-PC-BM-2), the crisis informatics case study provides a nice example of how social media can be mined to provide relief faster and better allocation of resources in case of emergency situations. The healthcare case study also exemplifies the provision of better services by enabling more accurate and timely diagnosis, as well as efficient treatments. In the culture case, improving data access raises the visibility of collections and drives more traffic to cultural heritage sites.

With respect to more targeted services (E-PC-BM-3), personalized medicine is a prominent example in the healthcare case study through the detection of rare genetic disorders that may not otherwise attract the public attention. In the smart cities case study, there are different services that can be provided to meet the personalized needs of citizens, e.g. in the area of mobility.

Cost-effectiveness (E-PC-BM-4) is another benefit that can be reaped with big data. In the crisis informatics case study, better resource allocation for humanitarian organizations is reported. Similarly, the focus group of the smart cities case stressed the immense efficiency increase potential along the dimensions of time, costs and resources. Indeed, optimization of big data utilities through data analytics (E-OC-TEC-2) is also identified as a positive externality in the smart cities case.

Besides the aforementioned positive externalities, there are also some negative ones that can hinder operational efficiencies; namely, E-PC-BM-5 and E-OO-BM-7. The need of skills and resources (E-PC-BM-5) is a major concern for the research organizations that lead the crisis informatics and the healthcare case studies. This externality might go unnoticed by the cost-effectiveness that can be achieved with big data (see E-PC-BM-4 above). However its importance should not be neglected, especially within organizations with tiny budgets and/or without specialized IT personnel. Concerning the inefficiencies associated to big data (E-OO-BM-7), this externality reflects that the use of big data pushes increased demands in computing power, data storage or network capabilities, and may thus cause inefficiencies, e.g. due to excess of pre-computed data that is later on discarded.

Beyond the cases studies ran in BYTE, big data can be applied in almost every sector to increase efficiency – see for instance McKinsey's 2011 report[1] for a number of ways that industries and domains use big data to make things run more efficiently. For example, the eBay Big Data Analytics Programme[2] aims to provide better search and recommendation services to their customers (E-PC-BM-2 and E-PC-BM-3) by applying big data analytics to their massive datasets.

Public bodies are also beginning programmes to gain operational efficiencies. In this regard, the Australian Public Service Big Data Strategy[3] aims to improve decision-making, targeting the delivery of services, and thus productivity, which, in turn, can substantially reduce government administrative costs (E-PC-BM-2 and E-PC-BM-4). Another example is the United Nations Global Pulse Initiative[4] that pursues to use big data technologies and tools for humanitarian reasons, especially in developing areas (E-PC-BM-2 and E-PC-BM-3).

Interestingly, the negative externality E-PC-BM-5 has also been identified – the European Space Agency Big Data Initiative[5] acknowledges problems to make accessible multi-petabyte datasets or managing streams of ~100s of GBs per day. One way to diminish the negative consequences of E-PC-BM-5 is through public expenditure; for instance, the aforementioned Global Pulse Initiative that aims to lower the systemic barriers to big data for development adoption. Similarly, the European Bioinformatics Institute[6] provides infrastructure and training to scientists interested in bioinformatics; while the CERN Worldwide LHC Computing Grid have developed a distributed computing infrastructure to deal with the scale and complexity of data from CERN's Large Hadron Collider (LHC). Private organizations can also offer solutions to address externality E-PC-BM-5; in this respect, Teradata[7] is a commercial enterprise that provides the necessary tools and platforms for the effective implementation of big data initiatives.

2.1.2 Innovation

Besides improvements in operational efficiency, big data can also foster innovation, as illustrated by externalities E-PO-BM-1, E-PC-DAT-1, E-OC-DAT-1 and E-OC-DAT-2. In the BYTE case studies, we have observed significant positive effects associated with open data

[1] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A. H., "Big data: The Next Frontier for Innovation, Competition, and Productivity," McKinsey Global Institute, 2011.
[2] Saran, Cliff, "Case Study: How Big Data Powers the eBay Customer Journey", computerweekly.com, 29 April 2014 http://www.computerweekly.com/news/2240219736/Case-Study-How-big-data-powers-the-eBay-customer-journey
[3] Lohman, Tim, "Australia Wants to be World Leader in Big Data Analytics", *Zdnet,* 5 August 2013 http://www.zdnet.com/article/australia-wants-to-be-world-leader-in-big-data-analytics/
[4] UN Global Pulse, "White Paper: Big Data for Development: Opportunities & Challenges", May 2012. http://www.unglobalpulse.org/projects/BigDataforDevelopment
[5] Mazzetti, Paolo, Angelo Pio Rossi, Oliver Clements, Stefano Natali, Simone Mantovani, Maria Grazia Veratelli, Joachim Ungar, John Laxton, "Community Reporting,· D120.21 EarthServer Project, 31 August 2013.
[6] The European Bioinformatics Institute, "Home", 2015. http://www.ebi.ac.uk/
[7] Teradata, "Turning Data – Big and Small into Huge Business Opportunities", 2015 http://www.teradata.com/

practices and unlocking datasets. Specifically, new products and services based on open data (E-PO-BM-1) is exemplified in the environment case study: the availability of satellite data enables society to be creative and create added value by releasing services such as Climate Service Center 2.0.[8] We can find another example of E-PO-BM-1 in the cultural case study; as heritage institutions release open data, other stakeholders can create services such as guided tour models for tourism purposes, which prompt people to travel and view the original version of what they see online.

While e-government projects focused on operational efficiency, initiatives such as Open Government efforts sought to promote public service transparency, civic participation, and inter-departmental collaboration. This could be achieved by sharing public sector infrastructure, seamless information sharing with other agencies, pushing core competencies to improve service delivery and engaging external entities, such as academies and businesses.

Government data is an important source that can drive innovation (E-PC-DAT-1). In case of the digital city, investment by the public sector into the data infrastructure of a city and the subsequent opening of this infrastructure as a utility/commodity can create a win-win situation that will ignite new business and increase welfare. Indeed, the Norwegian regulator publishes petroleum open data as a driver for competition. This is especially important for small companies since collecting data is extremely difficult and expensive. Moreover, reporting obligations benefit the petroleum industry as a whole, avoiding companies to duplicate efforts on data collecting activities.

Other positive externalities in the innovation area include innovation through open data (E-OC-DAT-2) and enhancements in data-driven R&D through standardization (E-OC-DAT-1). In the former instance, the crisis informatics case study illustrates how open data from Twitter can be repurposed to provide help in emergency situations. In the latter, the environment case study reflects that data standardization is another important enabler for innovation.

Despite the positive impacts associated with open data reported above, there is also some reluctance to open data (E-OO-DAT-5). In the energy case study, oil & gas companies are generally hesitant to share their data, especially in case of commercially sensitive assets, as in petroleum exploration. However, there is an internal debate among operators about this position, and opening data is proposed to exploit added-value services. Similarly, some stakeholders in the smart cities case study reported no incentives for data sharing required at that big scale.

Looking beyond BYTE case studies, we have identified the externality E-PO-BM-1 in the LHC, EBI and ESA initiatives that promote innovation through the release of massive datasets as open data, corresponding to LHC fundamental particles experiments, bioinformatics and earth observations, respectively. Government data is published within the UK Data Service[9] and the Australian Public Service Big Data Strategy, thus supporting externality E-PC-DAT-1.

[8] Climate Service Center, “Home”, 2015 http://www.climate-service-center.de/
[9] UK Data Service, “About us”, 2015 http://ukdataservice.ac.uk/about-us

Besides, eBay has released its distributed analytics engine, Kylin[10], to promote innovation within the open source community (E-OC-DAT-2).

Finally, Global Pulse is using different ways to address the negative effects of externality E-OO-DAT-5. In this sense, a top task for the group is to persuade telecommunications operators, and the governments that regulate and sometimes own them, to release some of their data. Moreover, Global Pulse advocates the concept of "data philanthropy" and the creation of a public "data commons,"[11] in which companies contribute large customer data sets, stripped of personally identifying information, for research on development and public health.

In order to foster innovation from government data (E-PC-DAT-1), there is a need also to understand the potential role for government regulation. Government can offer facilitator role by providing the infrastructure and framework necessary for data-intensive businesses to prosper, but should not actively intervene with the details of how big data business models develop. It is vital that any regulation is intended to facilitate, rather than restrain innovation and growth in the sector. Moreover, given the growing level of expertise in industry, private-sector stakeholders should be involved in any new regulatory process.

2.1.3 New knowledge

Public insights can become an important source of new ideas for how to drive value for the people and to stay competitive. To sustain an advantage in a digital economy, businesses should develop an in-depth understanding of public they service and markets they operate in.

The application of analytics and data mining techniques to large datasets can serve to gather public insight – externality E-PC-TEC-1 – and generate new knowledge – externality E-PC-TEC-2. For instance, the culture case study illustrates the creation of value by cross-linking datasets (E-PC-TEC-2). Another example corresponds to the crisis informatics case in which the needs of citizens affected by a crisis can be predicted through improved situational awareness (E-PC-TEC-1).

With respect to non-BYTE initiatives, innovation programmes such as the US Big Data Research[12] and Development Initiative, Australian Public Service Big Data Strategy or Global Pulse provide funding opportunities to generate new knowledge. Scientific initiatives such as CERN's LHC, ESA, EBI or deCODE[13] also exemplify the potential of big data for research –

[10] eBay, "Announcing Kylin: Extreme OLAP Engine for Big Data", 20 October 2014. http://www.ebaytechblog.com/2014/10/20/announcing-kylin-extreme-olap-engine-for-big-data/

[11] Pawelke, A. and Tatevossian, A.R. "Data Philanthropy: Where Are We Now?" United Nations Global Pulse, 2013. http://www.unglobalpulse.org/data-philanthropy-where-are-we-now

[12] Office of Science and Technology Policy, "Obama administration unveils "big data" initiative: Announces $200 million in new R&D investments", Press release, 29 March 2012, p. 1. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf

[13] deCODE Website - Genetics, "Unrivaled Capabilities". http://www.decode.com/research/

as an example, the Higgs Boson was discovered using the big data infrastructure and data from CERN's LHC.[14]

In most industry sectors, analysis and visualization of big data sets require development and integration of innovative algorithms with the best interactive techniques and interfaces. Because there is the need to simultaneously manage both, big data and large amounts of computation, to analyse the data. This complexity continues to increase rapidly. Thus, the workflow should be able to evidently support decisions such as when to move data to computation or computation to data.

2.1.4 Business models

With the emergence of big data, we can find innovative business models (E-OO-BM-2, E-PO-BM-2) and community building and innovation (E-OO-BM-1) externalities to complement revenue streams from existing products, and to create additional revenue from entirely new (data) products. In the healthcare case study those innovative business models are quite apparent: there are laboratories specialized in genomic sequencing and the healthcare industry is moving to the development of marketable treatments and therapies. In the energy case study, we can find companies based on data generation, e.g. seismic surveys, as well as organizations specialized in data analytics, such as well integrity. Moreover, there are ongoing partnerships around data (E-OO-BM-1) in which one party typically provides the data, while another performs some analytics. Condition-based maintenance is a common example of this kind of collaboration, in which equipment is instrumented to collect data and analytics are then applied for early detection of potential failures before they occur and to improve the efficiency of equipment.

While the previous examples show positive economic impacts associated with business models, there are a number of negative externalities that need to be addressed. Indeed, lack of adequate big data business models (E-OO-BM-8) was found in 5 of the case studies; for instance, in the culture case study private organizations have not yet found appropriate ways to monetize culture data or to generate commercially-viable applications. Similar arguments were used in the smart cities case study, stressing the uncertainty of data markets. Indeed, we found in the energy case study that big data technology is ready to make significant changes in areas like automated drilling, but return of investments were not assured with existing business models.

The introduction of big data can also challenge traditional non-digital services (E-OO-BM-3). For example, the smart cities case study found this risk that is likely to affect companies relying on old ways of procurement. In the crisis informatics case, the privatization of essential utilities by large-scale corporations (E-OC-BM-8) was stressed: this may result in resources provided by tax payers and philanthropists to humanitarian organisations ultimately being used to benefit large technology providers and other companies. In the smart cities case, small organizations regret their competitive disadvantage with respect to

[14] CERN, "CERN experiments observe particle consistent with long-sought Higgs boson", Press Release, 4 July 2012. http://press.web.cern.ch/press-releases/2012/07/cern-experiments-observe-particle-consistent-long-sought-higgs-boson

large corporations (E-OO-BM-5); their argument is that big contractors have the scale and resources to create a working solution for one city and sell it to the next with minimal or no changes.

The environment case study found additional negative externalities linked with business models. The digital divide (E-OO-DAT-1) was motivated by the fear of existing big corporations, the rising inequality caused by elite skills in data and the unequal opportunities in data access for SMEs. Further, reduced market competition (E-OC-BM-7) is observed with the emergence of a few dominant players that drive out competition and create their own data silos. Besides, public environment institutions complain that open data puts the private sector at a competitive advantage (E-PO-DAT-1) since they do not have to open their data, but have access to public data for free.

With respect to the non-BYTE initiatives, eBay and Teradata are two examples of successful business models in big data (E-OO-BM-2): eBay is using the accumulated data to achieve efficiency gains, while Teradata is a technology provider to facilitate the adoption of big data. Public agencies can help to diminish some of the negative effects associated with business models. In this sense, the US Big Data Research and Development Initiative aims to fuel start-ups[15] and thus provide broader economic gain (E-OO-BM-5). Global Pulse also aims to reduce the digital divide in developing countries (E-OO-DAT-1). In addition, the aforementioned public agencies aim to spread cross-sector and cross-industry collaborations (E-OO-BM-1) to increase the chances of positive impacts associated with big data.

As a whole, big data alone does not guarantee commercial success. In fact, when considering the success of businesses in general, a higher probability exists for failure. For instance, start-up companies often have interesting ideas for using big data, for example in mobile applications to identify traffic hotspots, or as ratings systems for local restaurants, but fail to develop plans for profitability and sustainability, such as using the ratings system to drive a service like Airbnb.[16] The likelihood of success increases when businesses are structured around innovative business model along with clear goal setting that identifies the intent of data to drive real-time decisions, for example, Uber, Airbnb, and Amazon.

2.1.5 Employment

The need of data scientists and data analysts was identified in three of the BYTE case studies (E-OC-BM-3). In the smart cities one, there was also a reflection about employment losses for certain job categories that will be replaced by technology (E-OC-BM-5).

[15] National Science Foundation, "NSF advances national efforts enabling data-driven discovery", Press Release 13-188, 12 November 2013.
http://www.nsf.gov/news/news_summ.jsp?cntn_id=129244
[16] https://www.airbnb.com/

Employment market analyses commonly mention the scarcity of data scientists.[17] To address this need, public programmes such as the US Big Data Research or the Australian Public Service Big Data Strategy include action points to improve education and training to meet the increased demand.

Moreover, earlier workflow models for data-intensive industry allowed for a separation of concerns between computation and analytics that is no longer possible as computation and data analysis become more firmly interlinked.

In recent day, a new kind of analytics technology focuses on empowering an organization's in-house talent to manage analytics directly on Hadoop, giving it data science capabilities without getting caught in the data science talent hunt. For example, the Actian Analytics Platform–Hadoop SQL Edition[18] allows SQL programmers to connect and analyse data natively in Hadoop.

2.2 OVERVIEW OF DATA SOURCES MAPPED FROM THE CASE STUDIES

This section summarizes the data sources that were mapped in the case studies. Mapping the data sources is an aid in the mapping of economic activities in section 0.

One interesting observation from the case studies is that the "big data" related economic activities produce and consume a large number of different data sources, and in terms of the "V's", there is a large number of disparate sources to be integrated, and consequently, a large "Variety" in the data types produced and consumed. The crisis study (as well as the energy case study) exemplifies this aspect. It uses many types of data from many sources to improve the situational awareness to inform decision-making. The data are not limited to numeric data that can be stored and analysed in spreadsheet applications. On the contrary, the data to be handled and analysed include text data from Twitter, text data from news media, aerial images from aircraft and satellites, geospatial data like maps of rivers and roads, time series data from sensors, and more. One important conclusion then, is that the majority of this data does not fit into spreadsheets, and cannot be analysed with traditional statistical data analysis tools.

The analysis issues in the context of making a coherent situational awareness seem to bear resemblance to issues in the "sensor fusion" field in autonomous vehicles where the objective is to harmonise sensor signals from various sensors (e.g. cameras, lidar, radar, audio) to make a coherent map of obstacles in the vicinity of the vehicle.

Table 2. . Main data sources employed in the crisis informatics case study

Data source
Geo-localized tweets (from Twitter)
Text messages

[17] Thomas H. Davenport, D.J. Patil, "Data Scientist: The Sexiest Job of the 21st Century", *Harvard Business Review*, October 2012. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

[18] http://esd.actian.com/Express/readme_HSE_2.0.html

News media
Aerial and satellite imagery
International humanitarian organisation data

Table 3. Main data sources employed in the culture case study

Data source
Cultural metadata (Linked Open Data)
Cultural works (digital images, sound recordings, texts and manuscripts)

Table 4. Main data sources employed in the energy case study

Data source
Seismic data
Geology models
Petroleum production data
Top-side, subsea and in-well sensor data
Drilling data
Knowledge repositories
Reference datasets (licenses, fields, wellbores, discoveries, operators and facilities)

Table 5. Main data sources employed in the environment case study

Data source
Space component (satellite data)
In-situ component (sensor data, e.g. rain gauges)
Service component (models)
Utilities & infrastructure data
Historical and archaeological data
Government agencies data (demographics, etc.)
Social media

Table 6. Main data sources employed in the healthcare case study

Data source
Blood samples
DNA samples
Sequenced genetic data
Reference sequence
Medical repositories (genomic databases)
Electronic health records (EHRs)

In addition to the data mapped in the healthcare case study, there are typically additional data produced and consumed in patient care related to diagnosis and treatment, including imaging, molecular, cellular, electrophysiological, chemical, patient behaviour and sentiment data, epidemiological, x-rays, research articles, insurance claims, wearables data, medication history, social media and pharmaceutical R&D data to name a few.

Table 7. Main data sources employed in the smart cities case study

Data source
Mobility data, e.g. location of buses
Utilities usage and planning
Sensor devices data (from smart buildings, lighting systems, etc.)
Sensor stations data (from train stations, airports, city centre, university)
Citizen data (lifestyle, taxes, etc.)
Mobile phone data

2.3 Overview of economic activities mapped from the case studies

This section summarizes the "big data" related economic activities that were mapped in the case studies. Big data related economic activities are activities that generate or consume "big data". As the aim of this report is to suggest how to capture the benefits and reduce the costs and negative impacts of big data related economic activities, it is helpful to summarise the activities first.

An interesting observation from the case studies is that many of the "big data" activities related to data production are related to **integrating or linking data** from disparate sources. The culture case study also highlights the aspect of **producing metadata** to make the data usable for consumption, while taking into account multilinguality challenges. **Classification of data** is also a central activity, exemplified in the crisis informatics (classifying tweets) and healthcare (classifying DNA variants) case studies. **Prediction activities** are also applied in some of the case studies, e.g. weather forecasting or discovery of petroleum deposits. Furthermore, **decision support activities** can also be found in mobility planning (smart cities) or drilling operations monitoring (energy).

Table 8. Big data related economic activities – crisis case study

ID	Activity description
CR1	Crisis mapping by mining and automatically classifying tweets from Twitter
CR2	Predicting needs for citizens affected by crisis
CR3	Crowdsourcing the collection of imagery and dissemination on social media
	Analysis of photos and messages on social media
CR4	Linking tweets with social media photos, text, and satellite photos to identify infrastructure damage
CR5	Crisis projects develop free and open source computing services (not data but SW)

Table 9. Big data related economic activities – healthcare case study

ID	Activity description
HE1	DNA sequencing (=digitizing genes); DNA variant detection and classification; DNA data curation
HE2	Digitizing clinical data (=make EHRs)

HE3	Digitizing prescriptions (eResept in Norway)
HE4	Linking and integrating health data: DNA, EHRs (=clinical), imaging, molecular, cellular, electrophysiological, chemical, patient behaviour and sentiment data, epidemiological, x-rays, research articles, insurance claims, wearables data, medication history, social media, pharmaceutical R&D data , etc.
HE5	Publishing genome repositories as open data[19]

Table 10. Big data related economic activities – culture case study

ID	Activity description
CU1	Generation of cultural metadata
CU2	Interlinking of cultural digital objects
CU3	Improving access and visibility of cultural heritage works through the PECHO portal
CU4	Reuse of cultural metadata in subsequent activities, e.g. tourism guides

Table 11. Big data related economic activities – energy case study

ID	Activity description
EN1	Discovery of petroleum repositories
EN2	Reservoir monitoring (permanent equipment in some cases)
EN3	Monitoring drilling operations
EN4	Monitoring well integrity
EN5	Improving the efficiency of equipment (condition-based maintenance)
EN6	Designing new data-driven products, e.g. subsea compressors
EN7	Improving safety and environment surveillance

Table 12. Big data related economic activities – environment case study

ID	Activity description
EV1	Prediction (e.g. climate change or weather forecast)
EV2	Decision support (e.g. crisis management)
EV3	Control systems (e.g. traffic)
EV4	Domain-specific (farming, tourism, food industry…)
EV5	Civil usages, e.g. virtual representations of the planet

Table 13. Big data related economic activities – smart cities case study

ID	Activity description
SC1	Reporting city activities
SC2	Mega and micro event management
SC3	Understanding citizen behaviour, e.g. mobility usage patterns or health habits
SC4	Optimized city planning
SC5	Quantification of city projects through the use of sensors
SC6	Mobility management

[19] NIH. "1000 Genomes Project data available on Amazon Cloud", March 2012. http://www.nih.gov/news/health/mar2012/nhgri-29.htm

SC7	Emergency situations management

2.4 OVERVIEW OF BEST PRACTICES – CAPTURING THE BENEFITS

To compete in a digital economy, it has become obvious that both public and private firms must leverage their information assets. Thus, make the most of the *value* of information assets is one strategic objective for many businesses and organizations. Apart from Internet powerhouses that have effectively established information-driven business models, businesses in other sectors are in the early stages of exploring how to benefit from their growing pile of data, and put this data to good use[20].

After assessing the main economic externalities associated with the use of big data, we have analysed the BYTE case studies and other big data initiatives in order to identify ways to diminish the negative effects and to amplify the positive impacts. As a result of this analysis, we propose a set of 8 best practices that are summarized in Table 14. For each one we include the main actors that can put them into practice.

Table 14: Summary of best practices to amplify the positive economic impacts and to diminish the negative ones

Code	Best Practice	Suitable for			
		SMEs	Large corporations	Research institutions	Public agencies
BP-1	Public investments in infrastructures				X
BP-2	Funding programmes for big data				X
BP-3	Releasing open government data				X
BP-4	Persuading "big actors" to release some of their data		X		X
BP-5	Promoting big data in education policies				X
BP-6	Seeking opportunities in new business models	X	X	X	
BP-7	Looking for interesting data sources	X	X		
BP-8	Creating partnerships around data	X	X	X	X

2.4.1 BP-1 Public investments in infrastructures

Public agencies have many alternatives to drive the adoption of big data and diminish the negative economic externalities associated with big data. Funding big data infrastructures (BP-1) is one way to bootstrap the adoption of big data, especially in such cases that require substantial investments. The environment case study serves to exemplify this best practice; public investments in the aerospace industry make it possible to use satellite and sensor data for a myriad of applications ranging from weather forecasting to agriculture. Similarly,

[20] "Big Data Survey Europe", BARC-Institute, February 2013

the LHC computing grid is a distributed storage and analysis infrastructure that was designed to handle the massive volumes of data generated by the LHC.

The smart cities case study identified the need of data infrastructures for handling the acquisition, curation, storage and analytics. Given the scale and complexity of cities, a data platform that can be shared by the different stakeholders is preferable in the long run. Instead of monopolistic structures, open source platforms may be the answer for the digital city. The public sector can fund the investment into the data infrastructure of a city, and the subsequent opening of this infrastructure as a utility/commodity. The rationale is the creation of a win-win situation for all that will ignite new business and increase welfare.

2.4.2 BP-2 Funding programmes for big data

While big data infrastructures may be necessary in some specific cases, public bodies can also dedicate resources to fund research and innovation programmes in big data (BP-2). By doing this, it is possible to promote positive innovation and knowledge creation externalities – see the culture case study for an example. Moreover, these best practices can diminish the need of skills and resources (E-PC-BM-5) and some negative externalities associated with business models, e.g. competitive disadvantage of SMEs (E-OO-BM-5) or the digital divide (E-OO-DAT-1). Indeed, this is the aim of initiatives such as Global Pulse, the Australian Public Service Big Data Strategy or the US Big Data Research.

While the common goal is to harness the power of big data and analytics, specific programmes should develop their own vision, objectives and action points. In the case of Global Pulse, this initiative is about maximizing the insight gained for humanitarian causes, using big data technologies in novel ways to produce socio-economic benefits. The main goals are thus to create success cases for development innovation, to lower the systemic barriers to big data and to facilitate the creation of a big data development ecosystem. Earlier projects from Global Pulse were mainly proof of concepts aimed to obtain insight from live data, e.g. social media and online news. More recently, projects from 2012 are driven towards specific programme and policy related questions.

The vision of the Australian Public Service Big Data Strategy is to use big data analytics in the public sector to deliver service delivery reform, better public policy and protect citizens' privacy. The Strategy raises the issues of the value of data held by Commonwealth agencies and the responsibility to realise this value to benefit the Australian public, as well as the need to negotiate the privacy risks of linking, sharing and providing broader access to data. The implementation of this Strategy relies on the following six principles: data as a national asset to be shared by all; privacy by design; data integrity and transparency of processes; sharing of skills, resources and capabilities; collaboration with industry and academia; and enhancing open data.

In the case of the US Big Data Research and Development Initiative, the vision is to extract knowledge and insights from large data collections to solve some of the most pressing challenges of the US. The Big Data Senior Steering Group (BDSSG)[21] manages this initiative,

[21] https://www.nitrd.gov/nitrdgroups/index.php?title=Big_Data_%28BD_SSG%29

coordinating the work across the different agencies involved. Some action points are aimed to improve the tools and techniques for accessing, organizing and gleaning insight from available data, while there are specific plans for some domains: healthcare, environment, education, biology and genetics, defence, and intelligence gathering

Interestingly, the Industrie 4.0 Working Group recommends the establishment of demonstrators and flagship projects to drive the adoption of Industry 4.0 in Germany.[22] This recommendation is in line with best practices BP-1 and BP-2, serving to deploy and test existing solutions in different settings. Note that Industry 4.0 is closely related to big data, promoting the use of cyber-physical systems, Internet of Things and services in the manufacturing industry.[23]

2.4.3 BP-3 Releasing open government data & BP-4 Persuading "big actors" to release some of their data

The following two best practices (BP-3 and BP-4) are aimed to release new data sources that can thus be exploited. Specifically, open government data (BP-3) is a way to foster innovation, as exemplified in the energy and environment case studies. While governments have been gathering data on a mass scale for a long time, they have often been ineffective at using data.[24] The rationale of open government data is to encourage the private sector and society to extract their value. The UK Data Service is one prominent example of an open government data initiative.

Besides governments, large corporations are also amassing enormous datasets, although they are commonly reluctant to open their data (E-OO-DAT-5). Moreover, we have seen that SMEs and new business have difficulties accessing valuable datasets and thus are in competitive disadvantage with respect to large corporations (E-OO-BM-5). Therefore, there is a risk that access to data becomes a powerful barrier to market entry. Public agencies can try to diminish these negative externalities by persuading large corporations to release some of their data (BP-5), as in the Global Pulse initiative (see the concept of "data philanthropy"[25]); this can drive innovation and even serve to find new insight that can revert on society and the organization that released the data.

Corporate data sharing initiatives have been surveyed in this study[26], finding six different types of activities: (1) academic research partnerships in which a research institution gets access to a corporate dataset; (2) challenges in which applicants compete to develop new

---

[22] Kagermann, Henning, et al., *Recommendations for Implementing the Strategic Initiative INDUSTRIE 4.0: Securing the Future of German Manufacturing Industry*, Final Report of the Industrie 4.0 Working Group. 2013.

[23] Ibid.

[24] Mayer-Schönberger, Viktor, and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think*, Houghton Mifflin Harcourt, 2013. Ch. 6.

[25] Pawelke, A. and Tatevossian, A.R. "Data Philanthropy: Where Are We Now?" United Nations Global Pulse, 2013. http://www.unglobalpulse.org/data-philanthropy-where-are-we-now

[26] Verhulst, Stefaan. "Mapping the Next Frontier of Open Data: Corporate Data Sharing". United Nations Global Pulse. September 2014. http://www.unglobalpulse.org/mapping-corporate-data-sharing

applications or uses of data; (3) data sharing with trusted intermediaries, normally for commercial purposes; (4) providing access to application program interfaces (APIs) for testing, product development or data analytics; (5) intelligence products that provide insight about broad trends such as market conditions; and (6) pooling, in which corporations share their databases to conform an aggregated dataset.

However, corporate data sharing initiatives have been relatively scarce and for this reason measures for the open sharing of data should be devised. This can take the form of new regulatory frameworks and incentives to promote corporate data sharing – although what sorts of measures are best suited to promote corporate data sharing is still an open question.[27, 28]

2.4.4 BP-5 Promoting big data in education policies

The next best practice is geared toward the current scarcity of data scientists and data analysts. This need has been identified in the energy, environment and smart cities case studies (E-OC-BM-3). Moreover, big data studies and reports commonly report this demand.[29, 30] In this regard, public agencies can put in place education policies (BP-5); this is indeed one of the action items in the US agenda for big data,[31] the Industrie 4.0 Working Group[32] and the Australian public service big data strategy.[33] Some universities have already begun to offer courses or even programmes in big data or data science.[34] Besides, lifelong learning and vocational training offer many opportunities to get and improve the necessary skills in big data, e.g. through MOOCs,[35] courseware and workplace learning. As

[27] Ibid.

[28] Verhulst, S., Dunn, A., Samarajiva, R., Hilbert, M., Stempeck, M. "Corporate Data Sharing for the Public Good: Shades of Open" (panel recording). 3rd International Open Data Conference 2015. May 2015. http://opendatacon.org/webcast/recording-corporate-data-sharing-for-the-public-good-shades-of-open/

[29] Thomas H. Davenport, D.J. Patil, "Data Scientist: The Sexiest Job of the 21st Century", Harvard Business Review, October 2012. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

[30] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A. H., "Big data: The Next Frontier for Innovation, Competition, and Productivity," McKinsey Global Institute, 2011.

[31] Networking and Information Technology R&D Program Big Data Senior Steering Group (BDSSG), "The National Big Data R&D Initiative: Vision and Actions to be Taken", Predecisional draft, 26 September 2014. https://www.nitrd.gov/nitrdgroups/images/0/09/Federal_BD_R&D_Thrusts_and_Priority_Themes.pdf

[32] Kagermann, Henning, et al., *Recommendations for Implementing the Strategic Initiative INDUSTRIE 4.0: Securing the Future of German Manufacturing Industry*, Final Report of the Industrie 4.0 Working Group. 2013.

[33] Lohman, Tim, "Australia Wants to be World Leader in Big Data Analytics", *Zdnet,* 5 August 2013. http://www.zdnet.com/australia-wants-to-be-world-leader-in-big-data-analytics-7000018963/

[34] Carnegie Mellon University, "Overview: Carnegie Mellon's Interdisciplinary Approach to Data Science" http://www.cmu.edu/graduate/data-science/

[35] Pappano, L. "The Year of the MOOC." *The New York Times* 2.12 (2012): 2012.

an example, the Academy Cube initiative[36] offers unemployed graduates in ICT the opportunity to obtain targeted qualifications and puts them in contact with employers.

As for data science programs, data science includes elements of computer science as well as statistical science. Stanford University has created a "Master of Science in Statistics: Data Science" within their Department of Statistics whereas the University of California Berkeley has created a "Master of Information and Data Science (MIDS)" at their School of Information. Both curricula include courses in statistics, programming, and in particular distributed computing. The statistics courses typically provide courses in open source statistical software like "R", and the distributed computing courses are on open source platforms like the Hadoop ecosystem where distributed solutions like Spark are central. Last but not least, we observe the need for skills not only in data analytics but also high-performance computing (HPC).

In summary then, we see an evolution where the statisticians, econometricians, and economists are taught programming and distributed computing, whereas the computer science student is taught topics in statistics. The need for skills in distributed computing comes from the "Volume" aspect of big data. Large volumes of data must be managed on clusters of cheap hardware rather than on a monolithic supercomputer that is much more expensive.

Furthermore, we observe from the BYTE case studies that the future data scientist much be capable of handling a much larger "Variety" of data types than what can be stored in a spreadsheet. In particular, the crisis case study demonstrates this.

Finally, computationally heavy jobs, and the need for real-time processing, the "Velocity" aspect of big data, drives the need for fast execution of software programs. These fast executing programs are typically developed in compiled languages like "C" and "FORTRAN" rather than interpreted script languages typically used in the Hadoop ecosystem (e.g. Python, Scala).

2.4.5 BP-6 Seeking opportunities in new business models

This section highlights some business opportunities as well as new business models. The remaining three best practices that have been identified are targeted toward adopting a big data mind-set. While we have acknowledged challenges with the business models in our research, there are actually new opportunities to explore in this area (BP-6). In the healthcare and energy case studies, we have found some companies specialized in data acquisition and data analysis. Other examples include the provision of technology and infrastructure, as in the case of Teradata. Public sector organisations and large corporations can initiate an agile organizational structure that can adapt and change quickly. This includes leveraging server virtualization and other cloud solutions. In Telefónica, the mobile network operator is now driving business innovation outside its core business units and brands. As part of Telefónica Digital, the "Dynamic Insights" initiative has commercialized the analysis of movement data, creating incremental revenue from retail, property, leisure,

[36] http://www.academy-cube.com/

and media customers.[37] Other carriers have developed similar offerings, such as Verizon's "Precision Market Insights" service.[38]

Emerging data-driven data models should be focused on the generation, aggregation and analytics of data – see the taxonomy developed at the University of Cambridge.[39] With respect to the generation of data, some studies discuss the commercialization of Linked Data.[40,41] One of the conclusions is that easily substitutable data assets, e.g. some kinds of metadata, are difficult to directly monetize, but can be used to trigger indirect revenues such as traffic generation. On the other hand, data assets that are unique or difficult to generate carry the highest potential for direct revenues. Direct revenue models include subsidies, licensing, subscriptions and advertising.

Capturing opportunities for Europe in the DNA sequencing business

From the healthcare case study, it appears that there exist no competitive sequencing labs in Europe since the samples are sequenced in The Far East. It seems thus to be a market opportunity for *establishing European sequencing labs* and perform *increased research on sequencing technology* to offer more competitive DNA sequencing services.

Reducing the costs of linking digital healthcare data

There are many challenges in linking digital data. They include technical challenges as well as legal challenges. Addressing the technical issues, the linking effort could be made more cost-effective by applying *data exchange standards* for data formats, taxonomies, metadata, etc.

Reducing the costs of accessing heterogeneous oil & gas data

Similarly, the benefits of accessing and analysing the linked data could be amplified by improved data access technologies using ontologies.[42,43]

Providing domain experts with flexible access to big data is a major bottleneck in data-intensive industries like oil & gas. Current systems lack the flexibility to pose ad hoc queries that join needed data from different sources within a constrained time frame. In response to this need, the Optique platform offers an automated, end-to-end connection between information needs, expressed with familiar and comprehensible terms from an ontology,

---

[37] http://dynamicinsights.telefonica.com

[38] http://www.verizonenterprise.com/industry/retail/precision-market-insights

[39] Hartmann, P.M., Zaki, M., Feldmann, N. and Neely, A. "Big data for big business? A taxonomy of data-driven business models used by start-up firms". University of Cambridge. Cambridge Service Alliance. March 2014.

[40] Pellegrini, T., Dirschl, C. and Eck, K. "Asset Creation and Commercialization of Interlinked Data". LOD2 project deliverable 7.5, July 2014.

[41] Vafopoulos, M., "A framework for linked data business models." In *Proc. of the 15th Panhellenic Conference on Informatics (PCI 2011)*, Kastoria, Greece, October 2011.

and the underlying data sources.[42] A pilot study at Statoil has succeeded to integrate several huge databases with Optique, while petroleum experts were able to use the proposed search interface to author queries joining up to 8 relations.[43]

2.4.6 BP-7 Looking for interesting data sources

While we have included best practices for making more data available (BP-3 and BP-4), stakeholders should look for opportunities within existing datasets (BP-7). For example, the crisis informatics case study illustrates how available social data can be reused for humanitarian purposes without entailing the creation of an entirely new dataset. The cultural case study showcases an initiative for realising open data that can be employed to drive innovation and serve to create new products and services. Instead of creating their own resources, stakeholders should thus check existing sources of open data – for example, checking catalogues such as Datahub[44] – as well as commercial data providers like Socrata[45] or Factual[46].

2.4.7 BP-8 Creating partnerships around data

Finally, the creation of partnerships around data (BP-8) should be promoted to exploit the synergies of the different stakeholders. This way, organizations can contribute with their data, expertise or tools to drive innovation. We have seen this type of collaboration in the energy case study: there are pilots between oil companies and suppliers to conduct condition-based maintenance. Given the scale and complexity dimensions of big data, stakeholders should thus look for similar ways to cooperate and create this type of synergies. Moreover, the creation of data partnerships is one of the goals of the US Big Data Research and Development Initiative,[47] fostering multi-stakeholder partnerships similar to BYTE. In addition, public agencies can encourage re-use of other sectors best practices and standards to insure consumers' privacy and security rights, while enabling new services through agile access to data. In this regard, the Industry 4.0 recommends the sharing of best practices, particularly among SMEs, and the establishment of common user groups for technology users, manufacturers and trainers.[48]

---

[42] Giese, M, Soylu, A., Vega-Gorgojo, G., Waaler, A. et al. "Optique Zooming In on Big Data Access." *IEEE Computer*, March 2015.

[43] A. Soylu, E. Kharlamov, D. Zheleznyakov, E. Jimenez- Ruiz, M. Giese, I. Horrocks, Ontology-based Visual Query Formulation: An Industry Experience, in: *Proceedings of the 11th International Symposium on Visual Computing (ISVC 2015)*, Las Vegas, NV, USA, 2015.

[44] http://datahub.io/

[45] http://www.socrata.com/

[46] https://www.factual.com/

[47] Networking and Information Technology R&D Program Big Data Senior Steering Group (BDSSG), "The National Big Data R&D Initiative: Vision and Actions to be Taken", Predecisional draft, 26 September 2014. https://www.nitrd.gov/nitrdgroups/images/0/09/Federal_BD_R&D_Thrusts_and_Priority_Themes.pdf

[48] Kagermann, Henning, et al., *Recommendations for Implementing the Strategic Initiative INDUSTRIE 4.0: Securing the Future of German Manufacturing Industry*, Final Report of the Industrie 4.0 Working Group. 2013.

2.5 DEALING WITH EXTERNALITIES – SOME LESSONS FROM DIFFERENT SECTORS

This section maps some externalities from the case studies, and discusses potential strategies for dealing with them. This section then complements the previous "Best practices" section that dealt primarily with how to capture more of the benefits related to "big data" related economic activities whereas this section focuses on externalities as the term is defined by economists.

2.5.1 Externalities – a primer

External effects are the costs or benefits that affect a party who did not choose to incur that cost or benefit. Typical examples are pollution by factory near a river that has an impact on the population wanting to swim or fish in the river. The factory would produce more of their product than they would if they had to pay all environmental costs. On the other hand, positive external effects, like gardens and architecture will be less provided by private citizens, than would be the case if they received payment from neighbours for their benefit of the beauty.

Externalities are the reasons why governments impose *regulations, incentives, subsidies, and taxes*. Taxes include so-called Pigouvian taxes such as emission fees to compensate for the damage of pollution. There are four kinds of externalities: negative externalities on production, positive externalities on production, negative externalities on consumption, and positive externalities on consumption. There are negative externalities on production when production imposes cost on an unrelated party, e.g. pollution from a factory on all citizens. Similarly, there may be negative externalities in terms of privacy loss when datasets are linked, in that the linking activity deanonymises anonymous data. There are positive externalities on production when the production causes a benefit on an unrelated party. For example, if I make a beautiful garden, my neighbours benefit from the view, and the value of their property may increase. With big data, linking genome data with health records may contribute to improved diagnostics and treatments and thus better health in the population. Negative externalities on consumption occur when the consumption imposes a cost on an unrelated party, e.g. your cigarette smoking impacts me negatively. Positive externalities on consumption occur when the consumption imposes a benefit for an unrelated party. For example, if I vaccinate myself against swine flu, it is less likely that I infect you, so you may skip the vaccination.

In the next subsections, we provide some examples on how one has dealt with externalities in various sectors believing that the ideas could be relevant for other sectors.

2.5.2 Reducing the negative externalities of linking personal health care data

The most obvious negative externality linking patient data with DNA data and making central repositories of EHRs is the potential *loss of privacy*. This data will potentially be available to a much large number of health care workers, including health personnel that have no reason to access your EHRs.

One possible solution to address this is to have very strict authorisation systems. This solution has several drawbacks. For one, it is costly to maintain such a fine-grained authorisation system. The other issue is that in urgency one would like the health personnel to have access to your EHRs.

So, another solution is to opt for a "nudging" or incentive solution. One may provide negative incentives for those health care personnel that want to peek into your EHR by requiring that the person must log a reason for accessing your EHR and also send a notification to the owner of the EHR of the activity and who was the actor.

2.5.3 Providing negative incentives into peeking into citizens' income data

This way of dealing with this kind of negative externalities is already implemented by Norwegian tax authorities. In Norway, the income and wealth of all citizens were made publicly available some years ago. This lead to extensive peeking into your neighbours' income, and it had negative consequences for example for children of too poor or too rich parents.

The current solution is to provide a negative incentive for extensive peeking. Firstly, you must log into your "My tax pages" site to view the income of others. Secondly, the person you peek into is notified that you did it and when. This has reduced "tax peeking porn" drastically. This also makes it impossible for foreign criminals to get information on who the rich and wealthy are and where they live.

2.5.4 Diminishing the potential adverse selection in health data production

We may ask who contributes to the data production in healthcare. There are a large number of data generation sources, from genome data to electronic health records (EHR) to wearables.

Many of the examples address the externalities related to the "digital divide", the increasing inequality between the poor and the rich and between the well-informed and the less informed citizens, which is aggravated in the emerging "digital society".

Diminishing the biased selection of illnesses in genome research

A possible negative externality of genome data collection may be that the data are sampled from only the rich and wealthy part of the world. In the genome pilot diagnostic project at the Oslo University Hospital[49], the research focuses on inherited genetic disorders sampled from the Norwegian population. The focus is on *inherited disorders*. It is not clear that inherited disorders in Norway are representative of worldwide-inherited disorders. The Norwegian population arguably belongs to the wealthy minority of the world population. Thus, the biased selection of the data material means that there will be more research into

[49] Dag Undlien, "Implementation of diagnostic NGS services – Experiences and challenges", Norwegian Sequencing Centre, University of Oslo & Oslo University hospital, 2015.

the diseases of the *disorders of the wealthy*, at the expense of resources that could have been allocated to research benefitting the poor majority of the world.

How could one deal with this bias? Economic incentives are one option. The amount of allocated research funding could be based on the world frequency of various inherited disorders, i.e. based on the amount of people who would benefit in case a cure was found, and also based on estimates the cost of a potential cure, a cost benefit analysis in short. The problem is that the funding is from national sources, and therefore how to provide incentives to a national funder to fund research that contributes to the global population.

Diminishing the adverse selection in use of wearables

Who buys wearables like Fitbits and GPS cardio watches? We claim that a certain segment of the population buys these devices, the ones who are already fit and wealthy, rather than the overweight and poor. There are several reasons for this. One is the cost; a cardio and GPS wristwatch costs a few hundred euros. Activity monitors like Fitbit are more motivating for those who already are active and thus get positive feedback from the device, like from my cardio device: “excellent recovery”.

It is arguably a better investment in healthcare to collect data and monitor those who are likely to suffer illness due to lack of physical activity. Arguably, behaviour is main contributing factor to a healthy life or illness.

What are the options then to counter the bias in who wears wearables? One suggestion to offset this social inequality could be to *offer free or subsidised devices* to those in the risk zone, like the highly overweight. This could be a step in the direction of preventive health care. Other issues are that the data from the device must be of *medical quality* to be safe in monitoring those at risk, like the overweight. Furthermore, the device must *provide solutions to the problems* it identifies, providing advice and “nudging” into a less risky life style.

Diminishing large inequalities in health data insurance terms

Personalised medicine is the dream and promise of linking gene data with other health data like EHRs. A possible negative externality is too personalized and differentiated insurance premiums. A citizen with “bad” genes may risk worse insurance terms than a “good DNA” citizen. This is different from providing lower car insurance to careful drivers who are not involved in accidents. We want to give positive incentives to careful driving and a healthy lifestyle. However, we also want to take care of the poor and the sick in a human society by providing them with as equal opportunities as possible. An insurer basing the insurance terms on the “DNA risk profile” is therefore violating the insurance principle in society “all for one”. Personalisation in health-related issues could result in a lack of insurances for people with high risk factors, while people with little risk would pay much lower premiums. The end result on societal level could be a lack of finances for healthcare and a diminished healthcare for higher risk people.

On the other hand, more personalised insurance for issues linked with behaviour can have a positive effect by influencing this behaviour and therefore lead to less risky behaviour and costs on a societal level. Insurers would like to get hold of and use genetic research results to tailor insurance and attract the "best customers".

One option to counter the insurers' incentives is to regulate insurance policies and prevent insurers from diversifying the terms, i.e. "everybody pays the same". We acknowledge however that the general insurance principle is already violated in many areas. Young Norwegian car drivers pay a higher premium than the older drivers. For health insurance it is the other way around, the older pay a higher premium. Women are less involved in car accidents than men, but in this case, Norwegian law has forbidden insurers to offer lower insurance to women. On the other hand, teetotallers are granted lower insurance.

We are at present not able to offer any economic options like taxes and incentives to counter this externality. Rather, we think that the grounds on which data differentiation is allowed, will be a political and legal decision. These issues are therefore treated in more detail in section 4.4.2 on equality and discrimination.

2.5.5 Diminishing the "digital divide" related to governments publishing Open Data

This section addresses some potential externalities related to governments publishing Open Data. In Table 1, these externalities are categorised under "Innovation" and include:

E-PO-BM-1	New products and services based on open data
E-PC-DAT-1	Foster innovation from open government data
E-OC-DAT-2	Innovation through open data (and open source)
E-OC-DAT-1	Enhancements in data-driven R&D through standardization
E-OO-DAT-5	Reluctance to open data

Table 1 is focused on the opportunities for businesses to innovate from Open Data published by the authorities. However, we would like to bring the attention to potential negative effects for citizens. We also think that our suggestions on how to address the negative externalities for citizens partly are relevant for how to improve the positive effects for businesses.

In theory, the publishing of public Open Data is a benefit to all citizens. In practice, Open Data is more likely to increase, rather than reduce, the "digital divide" and social inequality. In practice, we argue that it is only open to a small elite of technical specialists who know how to interpret and use it, as well as to those that can afford to employ the specialists.[50] Often, "Open Data" in practice means access to an API (Application Programming Interface). Boyd and Crawford argue that you therefore often need a computational background.[51] In this respect, computer scientists are at an advantage compared to social scientists.

[50] T Roberts. "The problem with Open Data", *Computer Weekly*, 2012. http://www.computerweekly.com/opinion/The-problem-with-Open-Data

[51] D Boyd and K Crawford, "Six provocations for big data, A Decade in Internet Time", Symposium on the Dynamics of the Internet and Society, September 2011. http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=1926431

Furthermore, computer geeks tend to be males rather than females. In the end, who is asking the questions determines which questions are asked. Put together, Open Data can therefore potentially create a "digital divide" between male computer literates and the rest.

Another issue with Open Data is data quality and fitness for use. Noteworthy, the kind of data that is now being released from administrative systems as "Open Data" was previously collected or created for other purposes. It has undeniable potential value, but it also substantial risks for validity, relevance, and trust. As administrative data were created for other uses, they are therefore more difficult to use and interpret and therefore more subject to misunderstanding and misuse.[52]

How then address the bias in who benefits from Open data? One suggestion is to ensure that the data is *easy-to-use* also for those without a degree in data science. Governments must ensure that citizens become aware of this data resource, and they must provide free training.

A way of addressing low data quality and "fitness for use", governments must ensure that the released data are prepared. Metadata is crucial, and maintaining the metadata. Ontologies and tools on top of that like the Optique tool suite (refer to section 0) may further ease access into Open Data.

As for computer skills in education, there seems that women should be given strong incentives to pursue a computer science education, for example by grants, lower requirements for grades, etc. Furthermore, programming skills ought probably to be mandatory also in social sciences and business schools. To be extreme, "*programming literacy*" ought probably to be a future requirement for every citizen just as much as a "read and write literacy". See also section 2.4.4 for more suggestions.

2.5.6 Reducing data monopolisation

This section addresses some externalities related to competition and "competition disadvantage". In Table 1, these externalities are categorised under "Changing Business Models" and include:

E-OC-BM-7	Reduced market competition
E-OO-BM-5	Competitive disadvantage of newer businesses and SMEs

Monopolisation distorts competition and the marketplace. So with monopolisation of data, Boyd and Crawford claim that the "big data rich" inform and influence policy-makers.[53] The data owner can drive research agendas that suit their profit-seeking endeavours – and make it difficult for others to use data for other purposes.

[52] SS Dawes. "A realistic look at open data", 2012. http://www.w3.org/2012/06/pmod/pmod2012_submission_38.pdf

[53] D Boyd and K Crawford, "Six provocations for big data, A Decade in Internet Time", Symposium on the Dynamics of the Internet and Society, September 2011. http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=1926431

Google's monopoly and control of user data, of consumer behaviour data, that is, can be used to distort search results to favour its own services – specifically shopping. Digital services, however ubiquitous, seem less tangible and therefore do not appear so obvious a threat to commercial pluralism, innovation, and to consumer interests.[54]

A potential positive externality of Google's data monopoly is due to the value network effect: Larger audiences improve Google's data and make its products more accurate. A potential negative externality is that Google knows too much about consumers to the latter's disadvantage, driving up advertising rates with costs passed onto the consumers down the line[55].

How then diminish the negative externalities of the data monopoly of Google type companies? The Guardian[56] proposes to open up civic data on public services, infrastructure, roads and resources to start-ups and public organisations; the personal information gathered by search engines could be made available to researchers under strict ethical standards. Cultural products – books, music, film, news – could be mediated by digital public libraries.[57]

Other suggestions include applying regulations on "Antitrust" and execute actions that address Google's disproportionate control of user data. This could restore a competitive threat to Google and increase innovation in a range of sectors.

Regarding Google, The Huffington Post writes: "*From its illegal spying on Wi-Fi connections via its Street View program, which a 9th Circuit Court of appeals just this month confirmed violated the federal Wiretap law on a mass scale, to its pressure on manufacturers not to use rival geolocation services on Android phones to the illegal hacking of the Safari browser to collect user data, Google has expanded its control of user data not through customer choice but through illegal invasion of user privacy and abusive manoeuvres against corporate rivals."*[58]

Finally, to avoid the "digital divide" in the research community, i.e. the unequal access to big data from for example social media sites like Google and Facebook, it is possible to deny the publication of a research study unless all the data is also published. This has been advocated in the research field "Empirical Software Engineering", and in genome research under the

[54] The Guardian. "Google dominates search. But the real problem is its monopoly on data", *The Guardian*, http://www.theguardian.com/technology/2015/apr/19/google-dominates-search-real-problem-monopoly-data (published 19 April 2015)

[55] N Newman. "Taking on Google's Monopoly Means Regulating Its Control of User Data", *The Huffington Post*, http://www.huffingtonpost.com/nathan-newman/taking-on-googles-monopol_b_3980799.html (published 24 September 2013)

[56] The Guardian. "Google dominates search. But the real problem is its monopoly on data", *The Guardian,* http://www.theguardian.com/technology/2015/apr/19/google-dominates-search-real-problem-monopoly-data (published 19 April 2015)

[57] Ibid.

[58] N Newman. "Taking on Google's Monopoly Means Regulating Its Control of User Data", *The Huffington Post*, http://www.huffingtonpost.com/nathan-newman/taking-on-googles-monopol_b_3980799.html (published 24 September 2013)

"Bermuda Principles".[59] Without sharing the data, it is impossible neither to reproduce a study nor evaluate the claims.

2.5.7 Diminishing "Big mistakes" from correlational "empirical evidence only" science

This externality is linked to the externalities mapped in Table 1 under the categories "Operational efficiency" and "New knowledge" and include:

E-PC-BM-2	Better services
E-PC-BM-3	More targeted services
E-PC-BM-4	Cost-effectiveness of services
E-OC-TEC-2	Optimization of utilities through data analytics
E-PC-BM-5	Need of skills and resources
E-OO-BM-7	Inefficiencies associated to big data
E-PC-TEC-1	Gather public insight
E-PC-TEC-2	Accelerate scientific progress through big data

We argue first that there is a fundamental distinction between "efficiency" and "effectiveness". Providing faster healthcare treatment is efficiency; providing the right treatment is effectiveness. Efficiency can undoubtedly be improved by massive data and automated analytics and decision-making. A "Medical IBM Watson" cognitive machine can probably make better diagnostics than a human doctor in most cases because it can process massive quantities of data. However, there is also the danger that sometimes IBM Watson is really far off, but the machine is not aware of it, unlike an educated medical doctor who possesses "common sense" and can also in many cases better judge the quality and relevance of the data.

In a completely different field, software engineering cost estimation; Myrtveit and Stensrud found that experienced software project managers provided more accurate and more consistent cost estimates with the help of data analytics results than not-so experienced managers.[60] The latter relied too much on the output from the machine rather than adding their "expert knowledge" on top of it. In other words, the more experienced people could better judge the output from the analysis tools.

"Big data analytics" is often about finding correlations, "correlation analysis", not necessarily accompanied by any understanding of the phenomenon. The "new knowledge" acquired risks therefore being "far off, wrong knowledge".

When analysing consumer behaviour for the purpose of more targeted marketing, the correlation approach works, and is useful. It is not a catastrophe if you get an ad that is misplaced, or you do not get an ad that would have been relevant. In other contexts, "big

[59] NHGRI, "NHGRI Rapid Data Release Policy", National Human Genome Research Institute, 2013, http://www.genome.gov/10506376 accessed Oct 2015.
[60] Myrtveit, Ingunn, and Erik Stensrud. "A controlled experiment to assess the benefits of estimating with analogy and regression models." *Software Engineering, IEEE Transactions on* 25.4 (1999): 510-525.

data" empirical evidence without supporting theory can lead to wrong knowledge and thus big mistakes, if one assumes causality where there is only correlation.

How then address an exaggerated faith in the so-called "evidence-based" scientific approach? We suggest that there is a balance to be held between empirical knowledge and theoretical knowledge. "Theory" actually is about understanding the phenomenon. Without theory, we argue it is impossible to test hypotheses and draw inferences.[61] "Theory" or human expertise and sound judgment should not go out of fashion because we believe we can find all the answers in the data. An "intelligent" machine is actually not intelligent at all. It is not able to *judge* whether a conclusion it makes is sound or not. A machine – an automation – lacks *a priori* understanding of causality and therefore makes pure data-based, correlational-based decisions.

Related to the issue of "purely data-based decisions" is the issue of knowledge embedded in software and in complex algorithms, and therefore the knowledge might be with the computer professional rather than with the domain expert. As an example, the Norwegian tax rules are automated in a software system that calculates how much tax each citizen shall pay. One of the authors (Erik Stensrud) knows that the tax authorities on some occasions called the software developer to ask about how to interpret the rules. Only, the software developer knew in detail the interpretation of the tax rule, as he had to make interpretation decisions when translating natural language, which is imprecise, into a software algorithm, which is very precise.

One way to diminish these kinds of externalities is to educate domain experts in programming, i.e. "programming literacy", so that they can themselves develop the algorithms from a natural language specification. At a minimum, they should be able to read and understand the software algorithms.

2.5.8 Diminishing "Big mistakes" from "big" but biased, non-representative datasets

"Big data" is not equal to "representative data". In the crisis case study tweets were used a source for improving situational awareness. However, tweets typically are produced from mobile phones. Crawford[62] argues that we may get a skewed picture of the situation based on tweets. In the case of the "Sandy" storm on the US east coast, there were most tweets from central Manhattan. However, Sandy damaged more in other locations with a lower mobile phone concentration, and in addition, there were blackouts in these areas further biasing the tweet concentration. This would then potentially result in wrong conclusions with respect to the amount of damage and hence wrong situational awareness and finally wrong decisions with regard to resource allocation.

---

[61] Myrtveit, Ingunn, and Erik Stensrud. "Do arbitrary function approximators make sense as software prediction models?", *Proceedings 12th International Workshop on Software Technology and Engineering Practice (STEP)*, Chicago, IL, September 2004.

[62] K Crawford, MIT Technology Review, Youtube, 2013.
https://www.youtube.com/watch?v=JltwkXiBBTU

How then deal with this issue? The first step is to have additional information, such as the concentration levels of smart phones in various areas; and knowledge about where the blackouts are, preventing mobile phone communication. In short, using human judgment and common sense, and not just relying on the data as "the objective truth".

2.6 CONCLUSIONS

We have analysed the economic effects and externalities associated with the use of big data in the BYTE case studies, as well as in other relevant big data initiatives. The economic externalities treated are categorized into Operational Efficiency, Innovation, New Knowledge, Business Models, and Employment dimensions.

Improved operational efficiency is one of the main reasons for using big data. Big data is used to make better decisions, to optimise resource consumption, and to improve service quality and performance. Besides the aforementioned positive externalities, there are also some negative ones that can hinder operational efficiencies: the need of skills and resources and increased demands in computing power, data storage or network capabilities.

Big data can also foster innovation. Positive externalities in the innovation area include innovation through open data and enhancements in data-driven R&D through standardisation. Releasing open data is especially important for small companies since collecting data is typically difficult and expensive. Despite the positive impacts associated with open data, corporations are somewhat reluctant to open up, especially in case of commercially sensitive assets.

The application of analytics and data mining techniques to large datasets can serve to gather public insight and generate new knowledge – this was evidenced for example in the crisis informatics case in which the needs of citizens affected by a crisis can be predicted through improved situational awareness.

With the emergence of big data, we can find innovative business models and community building and innovation, while also the lack of adequate big data business models was raised as a negative externality. Further also the challenge to traditional non-digital services and the potential dominance by larger players was raised. Lastly, an employment concern is the need for data scientists and analysts, while also the threat of employment losses for certain job categories that will be replaced by technology was raised.

In order to diminish the negative economic impacts and to augment the positive effects, we have devised a set of 8 best practices. Public funding in big data infrastructures and innovation & research programmes are aimed to diminish the negative impact of scarce resources. Funding big data infrastructures is one way to bootstrap the adoption of big data, especially in such cases that require substantial investments. Public bodies can also dedicate resources to fund research and innovation programmes in big data.

Given the scarcity of data in some areas and the barriers of SMEs and start-ups with respect to data access, we have proposed two best practices aimed to release new data sources that can thus be exploited. This concerns government data but also corporate data, although it

remains difficult to develop proper incentives for corporate data sharing. Education policies have to address the current scarcity of data scientists and data analysts, but also promote the inclusion of data skills in a range of educational programs.

Finally, the last three best practices are geared towards a change in the mind-set and to boost the adoption of big data. Our case study research uncovered several opportunities for new business models, as well as opportunities for new uses of existing datasets. The creation of partnerships around data should be promoted to exploit the synergies of the different stakeholders.

While the proposed best practices focus on capturing the benefits of big data, we have also discussed specific strategies for dealing with the economic externalities in a number of sectors. This concerns providing negative incentives against a too broad use of personal data, like towards medical personnel checking medical records or citizens checking other citizen's income data made public by the tax administration. Another issue is addressing the digital divide, between rich and poor, informed and less-informed citizens or between older and younger generations. This can lead to adverse selection and biased data collection, while also the negative effects of personalised risk assessment in the insurance sector needs addressing.

Further externalities to address are data monopolisation and the risk for "big mistakes". The former distorts fair competition, while open data and public mediation can serve as countermeasures. With respect to "big mistakes", the risk is to over-rely on correlational "empirical evidence only" science or on biased, non-representative datasets.

3 SOCIAL AND ETHICAL EXTERNALITIES: CAPTURING BENEFITS FROM BIG DATA

3.1 SIMILAR BENEFITS AND SIMILAR SOLUTIONS FOR THE ECONOMIC AND THE SOCIAL DOMAIN

In the case studies several positive externalities were found resulting from big data practices. Positive social impacts of big data were improved efficiency and innovation. These reflect the similar positive economic externalities, but appear under social externalities when part of activities with social, non-economic aims. Other positive externalities were improved awareness, decision-making and participation. These are specific forms of improved efficiency and innovation in the social domain. They also appeared as political externalities when linked to political decision-making. In general these positive effects are similar across economic, social or political contexts. Therefore also the best practices discerned as part of the evaluation of economic externalities apply across those contexts. The analysis made in the previous chapter will not be repeated. Instead it will be further specified with an analysis of how big data changes interactions between actors.

3.2 EXTERNALITIES AND THE EFFECT OF BIG DATA ON INTERACTIONS

As mentioned earlier, the concept of externality is originally an economic concept and using it outside an economic context poses some difficulties. When considering non-economical externalities, we cannot make a purely economic evaluation of the effects on third parties. We will rather have to evaluate how different incommensurable objectives conflict with each other. The trade-off between the parties does not present itself in one (economic) dimension, but between several orthogonal dimensions, e.g. economic interests versus privacy.[63] What is a 'best' practice will depend on an evaluation of the importance of each of these dimensions, which is often very subjective. The result will rather be a set of best practices, which each present a different value judgment. Trade-offs are sometimes unavoidable. Still, these dimensions do not have to be considered in an inescapable trade-off, where improving on one dimension necessarily affects the other negatively. Often a solution can improve in several dimensions or without large negative effects on other dimensions. In these cases a best practice is one which can avoid negative effects, while having positive effects on some dimensions.

Big data practices lead to changes in interactions between different actors, be it individuals or organisations. The externalities uncovered in the earlier empirical work can in general be explained from these changes in interactions. Practices regulating interactions and which are aimed at preserving specific interests are affected by these changes, while also benefits resulting from big data practices can be explained from these interactions. To understand externalities and evaluate best practices to deal with them, we have therefore to start from the level of these interactions and analyse how big data affects these interactions. Based on

[63] We use deliberately the term 'orthogonal' and not opposing. Opposing implies a negative and positive pole on the same axis or dimension, and balancing between the poles is always a trade-off. On the contrary, an orthogonal axis in geometry enlarges the space within which solutions have to be sought with an extra dimension. This metaphor makes clear we are dealing with incommensurable objectives. The solution is rather one of maximizing the benefits or minimizing the negative impacts in several dimensions.

this analysis we can develop an account on how the practices managing these interactions are affected by big data. Similarly we can also explain the beneficial effects. With this theoretical framework we can further develop an in-depth evaluation of the externalities and the best practices to deal with them.

New developments in distributed computing like cloud technology, make it possible to deal with very large amounts of data at much higher speed. However, big data cannot be equated with these technologies or cannot be limited to these aspects of volume and velocity. It also implies qualitative changes in terms of what can be done with this data: a variety of structured and unstructured data sources can be linked much easier with each other and analysed in new ways. New business models are built upon the capacity to capture value from data through the development of a data value chain along which data is transformed into actionable knowledge. The incentive for big data is that it provides greater efficiency through more fine-grained and targeted control, and therefore also cost reductions. E.g. the condition-based maintenance of equipment (see oil & gas case study). It can also generate new visibilities of certain phenomena and makes possible new innovative business processes. The construction of a new data value chain implies a new way to optimize output or to create new products and services and generates new data flows between actors. It results in new or altered relations between these actors. New or augmented data flows present a much larger interaction between them. These data value chains have a central role in a data-driven knowledge economy and push organisations and administrations to open up their data sources and business processes in order to reap the benefits, resulting in a new 'data ecology' consisting of a diversity of actors providing, collecting or analysing data and acting upon the results.

The capacity to collect, process and analyse data on a much larger scale has turned data into a network good. The growing availability of data results in positive network effects. This means that when new data becomes available, also new value can be made by re-using older data sources in combination with these new data. The marginal cost of re-using a dataset is near-zero (except when augmented by other means, like the cost to obtain a license), while the new use can create value. A growing amount of data provides opportunities to make more value out of data by multiplying its uses or by making possible more fine-grained analysis. These positive network effects are an incentive to enlarge data collection. Therefore they also are an incentive for business models which are more intrusive towards potential data sources and which demand more interaction with other actors. E.g. a targeted advertising model becomes more efficient the more data it can collect and will thereby be an incentive for more intrusions of privacy.

These larger data flows provide a higher visibility of an actor and the capacity to process this data results in a more in-depth knowledge about him. Big data makes it possible to get a more fine-grained perception of an actor's behaviour, which can be used for faster and more targeted reactions towards this actor. This has an impact on organisational boundaries. These boundaries get penetrated by these data flows. They are much more difficult to uphold and often get different functions in this context of much higher interaction. Boundaries are a mean with which an actor maintains its identity and autonomy. From this identity follows an actor's interests and its capacity for autonomous action, or agency, is important to be able to pursue and protect these interests. Legal means

like property rights, organisational structures mapping which activities are done by whom within an organisation and who has access to which information, or technical means like the walls of a house define boundaries between the outside environment and the inside of an organisation or the private sphere of an individual. An actor uses such means to filter what is seen and known about him by the outside world and to keep others out. Impacts on these boundaries also impact the processes through which an actor preserves this identity and autonomy. On the boundary of an organisation gatekeepers or gatekeeping functions are situated, which open and close channels of communication and information flows.[64] The growing amount and changing characteristics of interactions means that such gatekeepers can become dysfunctional, their role has to be changed or new gatekeeping functions have to be found and implemented. Big data makes an organisational system more open and therefore gatekeeping becomes more complex. It enlarges interdependencies, which can be reciprocal but also very unbalanced, and opens an actor to outside power and control. Pfeffer and Salancik conclude that organisational boundaries are situated "where the discretion of an organisation to control an activity is less than the discretion of another organisation or individual to control that activity".[65] Discretion refers here to the capacity to make decisions and control resources. Changes in interactions do affect this capacity and makes organisational boundaries fuzzy. Organisations develop strategies to keep external uncertainty and dependence under control.[66] One such strategy is to limit this dependency and the resulting interactions. This clearly conflicts with the augmented interaction linked with big data-related business processes. Regulatory and protective practices reflect current gatekeeping practices and are therefore also vulnerable for the impact of big data. They have to adapt to remain effective when the organisational boundaries are continuously penetrated with data flows.

Data-driven business processes lead to a shift in the structure of transactions from an exchange of goods towards a delivery of services. An exchange of goods involves a minimal or limited amount of information exchange between actors, while the whole transaction takes place during a specific moment or limited time interval. However, the delivery of services generally is more continuous or regularly in time and stretches over longer periods. It often also involves a flow of information during the whole period of service delivery. We can notice such shifts linked to big data-related processes in the case studies. E.g. while earlier acquiring equipment consisted in a sale with no or limited information exchange after the sale, the practice of condition-based maintenance of equipment turns this acquiring of equipment into acquiring also a service. Data-driven business processes rely on continuous or regular data collection and monitoring activities and data flows are therefore not limited in time. This change in transactions has an important impact on the regulatory practices surrounding these transactions. Where the momentary or time-bounded character of a sale naturally limited the data flow involved and therefore the need for such practices, they become much more important when the data flow becomes continuous and have to

[64] Bekkers, Victor, "E-Government, Changing Jurisdictions and Boundary Management" in Victor Bekkers, Vincent Homburg, *Information Ecology of E-government. E-government As Institutional and Technological Innovation in Public Administration*, IOS Press, 2005, pp. 53-71.

[65] Pfeffer, J. and G.R. Salancik, *The external control organizations*, 1978, New York, Harper & Row, p.32, quoted in Bekkers, Victor, "E-Government, Changing Jurisdictions and Boundary Management", op. cit., 2005

[66] Bekkers, Victor, "E-Government, Changing Jurisdictions and Boundary Management", op. cit., 2005

adapt to remain effective. Maintaining organisational boundaries has to change from a time-bound control of a transaction towards the monitoring and regulation of continuous interactions.

Current regulatory practices still reflect this perspective of control of a limited amount of interactions. Often they are not able to scale towards more interactive and continuous data flows, due to the high transaction costs involved. E.g. the consent mechanism in data protection law and IPR presents a high transaction cost and prevents an aggregated treatment. Therefore it becomes an important barrier whenever a large amount of interactions takes places. More efficient and scalable mechanisms have to be adopted which reduce the transaction costs involved.

This general account of how big data practices affect interactions between actors we will now use to explain the positive and negative social and ethical externalities and evaluate best practices to deal with them.

3.3 POSITIVE SOCIAL AND ETHICAL EXTERNALITIES

3.3.1 Interoperability as key factor to capture benefits from big data

The positive impacts, as an externality or as an internal impact, can be understood as a result from these changing interactions. To clarify this we add some observations from Rob Kitchin to the analysis in the previous paragraph. In his book *The Data Revolution* Kitchin sums up some extra characteristics of big data aside the traditional 3 or 4 V-characteristics.[67] In particular he points to the larger resolution and indexicality of big data. The data collected becomes more fine-grained in resolution and it is aimed to be as detailed as possible. This is accompanied by the dataset becoming more indexical, meaning that objects and individuals can be more uniquely identified and singled out. Such larger indexicality also makes it easier to combine data, which leads to a further characteristic of data becoming more relational. Big data provides more possibilities, under the form of common or correlated data fields, to connect datasets and find relations between data and objects reflected in that data. This enables data to become networked. Further, big data systems are developed to be flexible, meaning that it becomes easier to extend the dataset with new data fields (extensionality) and to expand in size rapidly (scalability).[68]

The observed positive impacts of big data, like improved efficiency and innovation, improved awareness, decision-making and participation, can be related to these characteristics. The larger granularity makes more precise, targeted and individualized solutions possible, leading to improved efficiency. Similarly, larger granularity provides enlarged visibility and thereby improved awareness, which also enables decision-making to be more precise and more targeted. Further, the larger relationality and flexibility makes it

[67] The original 3V-definition by Gartner reflected both characteristics of big data and technical challenges big data techniques aim to solve. These extra characteristics do not have this character of technical challenges any more nor are they defining traits. Rather they are more general observations which according to Kitchin hold for most, but not all, big data systems.

[68] Kitchin, Rob, *The Data Revolution. Big Data, Open Data, Data Infrastructures and Their Consequences*, SAGE Publications Ltd, London, 2014.

possible to establish new data value chains leading to new innovative work processes. Participation in the case studies also reflected the larger granularity. In the smart city case study participation did stand for citizen-centric services with immediate feedback, while the crisis case study involved crowdsourcing to label its training data set. In both cases this is linked to the capacity to serve a wide range of individuals with a targeted response. In some of the case studies the positive externalities had less to do with the specific characteristics of big data, but more with the further resulting effect of making information easier available. In the culture and environment case study participation was brought up in a more general sense of making information more accessible. Decision-making, political or otherwise, can also be improved by becoming more evidence-based, as was raised in several case studies.

These extra characteristics formulated by Kitchin clarify how big data produces these positive effects. Big data's capacity to deal with a higher volume of data at a higher velocity makes it possible to develop business processes capable to deal with much larger interactions. With Kitchin's characteristics this account can be further specified: big data makes it possible to deal with a higher amount of interactions, such that **transaction costs get lowered while retaining the granularity** needed for more individualised and targeted responses. This lowering of transaction costs to process and analyse a large amount of data also makes producing fine-grained results possible. Their use leads to the positive impacts signalled in the case studies. Capturing the positive benefits therefore implies ensuring that no other obstacles result in a substantial rise of transaction costs or barrier, when not necessary for justified concerns. This translates into a focus on improving interoperability of datasets. A key element in the construction of data value chains is the interoperability of datasets, or assuring that datasets can be combined and analysed together. The efforts to be made to bring data together and to prepare it for combined analysis result in extra transaction costs. Lowering the need for these efforts by assuring the interoperability of datasets is an enabling factor to capture the benefits of big data processing. Improved interoperability of datasets turns this data into a network good and lead to positive externalities through positive network effects.

The European Interoperability Framework (EIF)[69] provides a useful conceptual model of interoperability levels: semantic, technical, organisational and legal interoperability. This conceptual model was developed for public services, but we use it here in a generalised meaning. Semantic interoperability concerns the meaning of data. It is about ensuring that this meaning is understood and preserved throughout the data exchanges. This involves metadata and vocabularies describing the meaning of data elements and their relations, or standardized formats of data. Improved semantic interoperability make it possible to link data sources with less effort and makes them more relational. Technical interoperability concerns the technical aspects of linking information systems and giving access to data, enabling the flow of data and its analysis. This concerns issues like providing web-based access, establishing data portals, etc. Organisational interoperability concerns how organisations cooperate in order to establish data value chains. Business processes have to be set up in such a way that the exchange of data is possible. Legal interoperability concerns

[69] European Commission, annex II of the "Communication: Towards interoperability for European public services" - COM(2010) 744 final

the (differences in) legal status of data and how to deal with them, like through licenses and data use agreements. Differences in legal status can pose a barrier to the linking of datasets or lead to limitations of data exchange. The legal and organisational interoperability levels leave no doubt about the fact that interoperability of datasets is more than just technique. It is (also) influenced by legal frameworks, organisational structures and needs investment in data quality in order to capture potential benefits.

These four interoperability levels also point to the different levels through which data flows can be regulated and where gatekeeping functions can be established. E.g. anonymisation limits semantic interoperability, while encryption limits access and can be seen as regulating organisational interoperability (by blocking semantic interoperability unless you have the key for decryption). Similarly the 'Terms of Use' of a data source can block or allow on the legal level its use for specific big data processing purposes. In the following treatment on mechanisms to deal with negative externalities we will therefore often encounter a mix of measures on the different levels. Such gatekeeping can be fully justified, but can also create unnecessary costs and barriers when it is done through mechanisms not adapted to the context of big data. E.g. the use of mechanisms based on individually negotiated transactions and authorisations can lead to large transaction costs and impede big data processing.

3.3.2 Examples of best practices

One important best practice in terms of interoperability is open data. Open data aims at making data sources available for new, undefined purposes and uses by other actors. Tim Berners-Lee, famous for developing the hypertext protocols, has formulated a 5-star scheme for open data.[70]
Open data is:

1) * available with open licence,
2) ** as machine-readable structured data,
3) *** in non-proprietary format,
4) **** using open standards,
5) ***** linked to other's data

This definition reflects the 4 interoperability levels. E.g. the technical and organisational level in the term 'available' (like through a data portal), the semantic level (machine-readable structured data, linked to other's data), the legal level (open licence, non-proprietary format, open standards).

Capturing the positive benefits of big data also implies designing business processes in such a way that they are compatible with the large amount of interactions, while including the necessary gatekeeping functions to prevent negative impacts. An interesting example of such effort can be found in the German Industrie 4.0-initiative. The term Industrie 4.0 points to a fourth industrial revolution, which after mechanisation, electrification and informatisation now concerns the introduction of the Internet-of-Things in industrial processes. Machines and sensors built into production and business processes get

[70] Tim Berners-Lee, "Is your Linked Open Data 5 Star?" (addition 2010), in: Tim Berners-Lee, "Linked Data", 18 June 2009, http://www.w3.org/DesignIssues/LinkedData.html

networked into cyber-physical systems, enabling the optimization of the production process and turning the whole into Smart Factories. In this context big data is an enabling technology among other technological components like the Internet-of-Things (IoT) or Internet-of-Services (IoS). The oil & gas case study already showed a current example of such automated industrial units, in this case at the seabed, packed with sensors and networked into large systems. This Industrie 4.0-initiative presented several design principles for these systems: Interoperability, Virtualization, Decentralization, Real-Time Capability, Service Orientation and Modularity. This example comes from the industrial sector, but these design principles remain important when developing large cyber-physical systems for non-economic or mixed purposes, e.g. smart cities. Interoperability points in this context to the ability of cyber-physical systems and humans to connect and communicate with each other. Technological enablers are the Internet of Things and the Internet of Services and the focus is most on technical interoperability. Virtualization points to the ability to create a virtual copy of the Smart Factory by linking sensor data with virtual plant models and simulation models. In this context big data plays an important role. However, this capacity to create a virtual copy does not mean that the system is fully centralized. A further principle is Decentralization, pointing to the ability of cyber-physical systems within Smart Factories to make decisions on their own. In the oil & gas-case study this came forward in the integration of data analytical and control capacity as a smart device into drilling equipment. Real-Time Capability points to the capability to collect and analyse data. Again big data plays an important role in connection with sensor-driven data collection. Service Orientation concerns the offering of services via the Internet of Services. An example encountered in the case studies is condition-based maintenance of equipment. More in general cloud services can entail several layers of services provided by different actors. This principle points to the shift to services linked to continuous data flows we presented earlier. Such multi-party environment is made possible by the last design principle Modularity. This makes a flexible adaptation possible by replacing or expanding units and services.[71] The large amount of sensors and the connectivity through the IoT creates a large amount of interactions, while other actors enter the factory through the service delivery via IoS and make the organisational boundaries fuzzy. This necessitates also a renewal of gatekeeping functions, as can be seen in the policy recommendations of the working group Industrie 4.0, which highlighted the need for security-by-design and new protections of corporate data, e.g. by a legal protection of trade secrets.[72]

In other words, capturing the benefits of big data demands improving interoperability on several levels, but also entails diminishing the fear for bearing the cost of negative impacts and the resulting distrust and reluctance to participate in big data practices. This can be done through restoring legal certainty and ensuring adequate balancing between several interests. For this we look at how to deal with negative externalities.

3.4 NEGATIVE SOCIAL AND ETHICAL EXTERNALITIES

[71] Hermann, Mario, Tobias Pentek, and Boris Otto. "Design Principles for Industrie 4.0 Scenarios: A Literature Review.", http://www.leorobotics.nl/sites/leorobotics.nl/files/bestanden/2015%20-%20Hermann%20Pentek%20%26%20Otto%20-%20Design%20Principles%20for%20Industrie%204%20Scenarios.pdf

[72] Forschungsunion and acatech, “Umsetzungsempfehlungen für das Zukunftsprojekt Industrie 4.0”, http://www.bmbf.de/pubRD/Umsetzungsempfehlungen_Industrie4_0.pdf

As noticed earlier the distinction between negative social and ethical externalities and legal externalities is often difficult to make, as the legal frameworks are meant to address the social and ethical problems. For coherence with the previous deliverables D4.1 and D2.1 we consider privacy in the chapter on legal externalities and equality and discrimination in this chapter on social and ethical externalities. But both issues turn up first as social and ethical externalities. The legal frameworks concerning these issues is one of the main instruments through which these externalities are addressed. Big data creates also legal externalities when it impacts on the functioning of these legal frameworks. Therefore we first give a general account on how the effects of big data on interactions between actors results in negative externalities. As this happens by disrupting the functioning of current legal frameworks, this account is applicable on both the social and ethical externalities considered in this section (equality and discrimination) and the legal externalities considered in the next chapter.

In this section on negative social and ethical externalities we will focus on equality and discrimination. Big data practices can have discriminatory effects which brings them in the realm of 'traditional' anti-discrimination legislation. Therefore we have to look at how the anti-discrimination legislation affects big data practices and can be applied on them. We will notice that it is less a question of adapting the legal framework, but rather of developing adequate tools to perform big data practices in compliance with anti-discrimination legislation. The data protection framework and the privacy-by-design approach will be an important source of inspiration for our recommendations on how to address the potential discriminatory effects of big data.

Another motivation for treating the effect of big data on equality and discrimination as a social and ethical externality is that it raises a lot of new ethical questions which are not addressed yet by the legal framework. These questions result from the fact that big data provides a much wider visibility of a lot of factors, different from the traditional anti-discrimination concerns like sex and race. It makes them available as a ground on which treatment can be differentiated. To which extent such differentiation is acceptable will need in the first place a lot of ethical and political debate, before the legal framework can be adapted to address these questions.

3.4.1 Consequences of changing interactions and appearance of negative externalities

The negative externalities which were uncovered in the case studies can be explained largely from the difficulties to adapt regulatory and protective mechanisms to the new characteristics of interactions.

The enlarged visibility and penetration of boundaries through big data appear as privacy problems for individuals. Also companies and other organisations have problems dealing with such visibility through growing data flows. They fear losing confidential information to competitors or becoming vulnerable to information leakage towards commercial partners, and becoming victim of rent-seeking by them based on this information. Power relations in value chains stretching over a range of organisations can be upset. Similarly it makes these

actors vulnerable to surveillance, discriminative treatment and political abuse. For both individuals and organisations this can lead to mistrust, uncertainty and a fear to lose control and autonomy. In such cases withdrawal and reluctance to participate in big data practices are a rational strategy to reduce uncertainty. Restoring certainty through other means, like adapted legal frameworks, would enable these actors to participate again and capture and contribute to the benefits of big data. The shift to services and the continuous interaction involved generate the need for new forms of gatekeeping and upholding boundaries.

The positive network effects of data also make protection techniques like anonymisation more vulnerable and unreliable on longer term. These network effects and its incentive to use more data can also lead to a propagation of discriminative effects present in data and result in unintended discrimination in other contexts. Decision-making can become more opaque when based on combining a wide range of data sources and make such decision-making less accountable.

The consent model in data protection law is based on a transaction model, allowing the use of personal data by others only after the consent by the data subject. This leads to large transaction costs, when a lot of data subjects are involved. Similar problems arise for other legal frameworks. Copyright and database rights make the use of data dependent on the consent of the author or the right holder. This leads to many difficulties to clear the rights when a lot of rights holders are involved, and by consequence to large transaction costs.

We can conclude that the negative social, ethical and legal externalities are a consequence of how big data changes interactions and of the inadequacy of legal frameworks to deal with these changes.

In the rest of this chapter this account will be applied on discrimination and used to evaluate best practices to deal with big data. The larger visibility of actors subjected to big data practices provides new grounds for discrimination. The improved awareness provided by big data, and the decision-making resulting from it, can be used for positive and negative purposes. Further, when data becomes a network good, old discriminative practices reflected in the data can proliferate and lead to unintended discrimination. Best practices to deal with discrimination have therefore to take these potential effects of big data into account. They also need to be effective in the context of larger interactions.

3.4.2 Equality and discrimination from a sociological and legal perspective

To discuss how big data can result in discriminatory practices, some important sociological and legal concepts concerning equality and discrimination need to be introduced.

The principle of equality is a long-standing ethical and legal principle. Linked to this principle is the prohibition of discrimination, which refers to the prejudiced or unjustified distinction in treatment of an individual based on their membership of a certain group or category. Important here is the term 'unjustified' as any decision-making about persons entails making a distinction in treatment. But certain grounds for doing so, like a physical or cultural trait, such as sex, ethnic origin, religion or sexual orientation, are considered unjustified. Legal protection against a broad range of discriminative practices has been developed,

although not everything which is perceived socially as a discriminative practice is covered by these prohibition. Anti-discrimination law and policy is an area which remains in full development.

Discrimination as a social phenomenon appears in many different outfits. In sociological literature three main mechanisms of discrimination are put forward: prejudice, rational discrimination, and unintentional discrimination. Prejudice concerns unfairly or unreasonably formed negative attitudes against a group, which can translate into stereotypes on which discriminative treatment is based. Rational discrimination does not start from negative attitudes towards a group, but from rational thinking where the lack of knowledge about an individual is compensated by using prior knowledge of the average performance found in the group to which the individual belongs. Again a form of stereotyping is used, although now a stereotype which can be justified on rational grounds, but which can still be considered unjustified on moral or legal grounds. E.g. when employers discriminate against women to avoid the cost of pregnancy leaves, when insurance companies take gender differences in behaviour or life expectancy into account to calculate insurance cost. Unintentional discrimination occurs when no intention to discriminate is present but still, due to the lack of awareness of the possible discriminatory effects, certain practices are used which lead to discriminatory results. E.g. the use of assessment tests which are culturally biased in employment procedures.[73]

This differentiation of discrimination is important as we will notice that anti-discrimination law depending on the discrimination ground provides different scopes of protection. EU anti-discrimination law consists of 2 elements.

First a general principle of equal treatment, which "requires that comparable situations must not be treated differently and different situations must not be treated in the same way unless such treatment is objectively justified".[74] But this principle remains weak and is only assessed by a marginal test.[75] A difference in treatment can be justified when it is based on "an objective and reasonable criterion".[76] When the difference in treatment is not arbitrary in the sense of being based on prejudice, but it can be justified by some rationality, no violation of this principle of equal treatment will be found. In other words, this general principle protects against prejudice, but not against rational discrimination.

Second, anti-discrimination law contains prohibitions of discrimination on specific protected grounds. The implementation of this prohibition is very varying depending on the ground. Article 21 of the EU Charter of Fundamental Rights contains a prohibition of "discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national

[73] Andrea Romei and Salvatore Ruggieri, "Discrimination Data Analysis: A Multi-disciplinary Bibliography", in Custers, Bart, Calders, Toon, Schermer, Bart, & Zarsky, Tal (Eds.), *Discrimination and privacy in the information society: Data mining and profiling in large databases*, Springer Science & Business Media, 2012, p. 110.
[74] EUCJ, Arcelor Atlantique and Lorraine and Others, C-127/07, 16 December 2008.
[75] Gellert et al., "A Comparative Analysis of Anti-Discrimination and Data Protection Legislations", in Custers, Bart, Calders, Toon, Schermer, Bart, & Zarsky, Tal (Eds.), op. cit., 2012.
[76] EUCJ, Arcelor Atlantique and Lorraine and Others, C-127/07, 16 December 2008.

minority, property, birth, disability, age or sexual orientation". Article 19(1) TFEU gives the EU the competence to "take appropriate action to combat discrimination based on sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation". Specific legal frameworks develop the protection against discrimination, although with varying scope.[77] The widest protection is provided against discrimination based on sex or race. The directives apply not only in the area of employment, but also concern access to goods and services and social protection. In these areas the directives also provide for institutional support under the form of equality bodies. Concerning discrimination based on religion, disability, age or sexual orientation, only a directive concerning employment applies, while no equality bodies are provided for (although national laws can extend the protection and provide for such equality bodies).[78] The European Commission has proposed in 2008 a Directive to extend this protection also to other areas, but this directive proved controversial for some member states and is still pending.[79]

Two important concepts present in all these implementations of the principle of equal treatment are direct and indirect discrimination. Direct discrimination is defined "to occur where one person is treated less favourably than another is, has been or would be treated in a comparable situation" based on one of the protected grounds. Indirect discrimination is defined "to occur where an apparently neutral provision, criterion or practice would put persons" having a particular protected feature "at a particular disadvantage compared with other persons, unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary".[80] The prohibitions of direct and indirect discrimination broadens the protection and prohibits also rational and unintentional discrimination. An example can be seen in the Test-Achats decision of the EUCJ.[81] This decision concerned the Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services. Concerning insurance contracts, it stated that sex could not be a factor leading to differences in premiums and benefits. However, article 5(2) allowed members states to make an exception and permitted differences

[77] Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin; Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation; Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services; Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast); Gellert et al., "A Comparative Analysis of Anti-Discrimination and Data Protection Legislations", op. cit.

[78] Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast)

[79] European Commission, Proposal for a Council Directive on implementing the principle of equal treatment between persons irrespective of religion or belief, disability, age or sexual orientation, COM(2008) 426 final, 2 July 2008

[80] Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin, art. 2(2); similar in the other anti-discrimination directives.

[81] EUCJ, Association Belge des Consommateurs Test-Achats and Others, C-236/09, 1 March 2011

"where the use of sex is a determining factor in the assessment of risk based on relevant and accurate actuarial and statistical data." But the EUCJ considered this exception not compatible with the objective of equal treatment and declared it invalid. In other words, the directive allowed in this specific case a form of rational discrimination, explicitly based on a protected ground. For the EUCJ this qualified as allowing space for prohibited direct discrimination. In the context of big data and data mining this implies that such forbidden grounds may not be explicit factors used in building a model to make decisions within the scope of the anti-discrimination directives. Similarly, using models which unintentionally discriminate on one of the protected grounds to make such decisions will be considered as indirect discrimination.

Now we have clarified the main sociological and legal notions concerning equality and discrimination, we can look at how big data can give rise to discriminatory practices.

3.4.3 Big data and discrimination

Big data practices can lead to specific forms of discrimination. In the case studies the risk for discrimination was raised as a concern in several case studies. It appeared in several forms. In the crisis informatics and the smart city case study it was raised as a concern for unintentional discrimination. E.g. the reliance on Twitter could lead to a bias against more disadvantaged groups where the availability of smartphones was lacking and the use of Twitter not widespread. In the environment and crisis case studies also the concern was raised that certain populations can become more visible or discernible, and therefore also more vulnerable for discriminatory behaviour or political persecution. This last concern for intentional discrimination will be treated under the political externalities. Intentional discrimination will only be treated here when the results of data mining and big data in itself reflect intentional discrimination.

Several sources of discrimination in data mining can be identified. These are in general forms of unintentional discrimination, although they can also mask intended discrimination. A first source of discrimination is the definition of the objectives of a certain data mining practice and which problem it is supposed to solve. This is not always an obvious task and the data miner has to translate these objectives into a formal question that can be processed by computers, like defining a target variable. This target variable is associated with a set of 'class labels' or possible outcomes. This can be discrete classes between which the algorithm has to classify the items (e.g. creditworthy or not) or a value to estimate (e.g. a risk score). This translation of a problem into a target variable and class labels can have the unintentional effect of negatively affecting certain groups. Such definitions of a target variable, in other words what is a 'good' result or what is a 'risk', can be very subjective and can be influenced by other factors, such as the need to translate it into measurable criteria. E.g. diplomas or academic degrees are measures of competence, but competences can be acquired in several other ways, like experience, self-study, and having a diploma can also reflect other social factors, like having the economic resources to pay for an academic study. This can also be varying over age groups, depending on how accessible formal academic or higher education was. The requirement of a diploma is therefore an easy measure for competence, but which can be biased and have unintended discriminating effects. Another example is using the prediction of tenure as a criterion for hiring decisions. When the data

reflects a shorter tenure for women due to pregnancy leaves or quitting of jobs to take care of children, this criterion can lead to an unintended but indirect discrimination of women. Such problems are of course not specific for data mining, but can get an extra and less visible dimension through the problem definition in data mining practices.[82]

A second source of discrimination can be the training data, or the examples used to build a model. Such model building relies on two assumptions. First that the characteristics of the population on which the model is build stay the same for the population on which the model will be applied in the future. Second, that the data is a reliable representation of the population. When these assumptions are violated, the model can fail and have unintended discriminatory effects due to forcing a model of a specific reality on a population outside this specific context. Further, the data can reflect old discriminatory practices and the model build on this data will exhibit the same discriminatory bias in its results. The first assumption does not hold when relevant factors reflected in the data change. This can vary from changing economic (e.g. income levels and employment rates in times of crisis or when the economy is booming) or other circumstances (e.g. changes in the availability of technology like smart phones). And when this first assumption holds, it can still be a problem when it reflects old discriminatory practices. The second assumption can also be problematic in 2 ways. The sample of the population can be an inaccurate representation of the actual population, or the selection of attributes can be an inaccurate or incomplete description of the sampled population.[83]

Certain parts of the population can be under- or over-represented in the data, which can skew the result in disadvantageous ways. This can be the unintentional result from the data collection method. For example, in the crisis case study the reliance on Twitter data was known to entail risks for biased data collection as it depends on the availability of Internet or smart phones. The lack of this technology can lead to dark zones where a disadvantaged population is not observed. Another example signalled by some authors is Street Bump in Boston and similar programmes, where road quality is checked through sudden changes signalled through accelerometers in smart phones. This leads to a bias in the data through under-representation of poor neighbourhoods, due to the lesser amount of cars and smart phones available to the population in these neighbourhoods. Such bias can also be the consequence of existing discriminatory practices. E.g. racial profiling by police can lead to an over-representation of certain population groups in criminal and security statistics.[84]

The selection of attributes to use as input variables, also called 'feature selection', can also have discriminatory implications for certain groups, if these groups are not well-represented by these attributes and the factors which do better account for the statistical variation within those groups are not present or at a too coarse granularity. Also in this case the second assumption is violated. The representation by a range of attributes is by definition a reductive presentation of a much more complex reality. It is impossible to take all factors

[82] Barocas, Solon and Selbst, Andrew D., "Big Data's Disparate Impact", 2014, available at SSRN 2477899 (2014).

[83] Calders, Toon, Indrė Žliobaitė, "Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures", in Custers, Bart, Calders, Toon, Schermer, Bart, & Zarsky, Tal (Eds.), op. cit., 2012.

[84] Ibid.

into account and this process of reducing a phenomenon to a range of attributes is therefore always a form of stereotyping. Such representation can fail to capture the relevant details needed to distinguish between the different outcomes, or to do so for specific groups. Attributes can be coarse proxies of certain phenomena and thereby lead to discriminatory outcomes. The controversial practice in the US of redlining is based on such a coarse proxy. Financial institutions used postcodes or geographical criteria as an easy accessible indication for creditworthy and refused loans or other services to people in specific areas. The postcode becomes also a measure of income or ability to repay loans. It is a cheap but very coarse indicator compared to an individual investigation of a person's creditworthiness, but also discriminates against people living in these areas. Similarly, but perhaps less obvious, other attributes can also have discriminatory effects.[85] Further, the attributes through which subjects are represented can be correlated. When attributes are independent they have an equal contribution to the decision-making in the model. When attributes are correlated, it is less clear to determine and control the contribution of attributes to the resulting model. On the other hand, not all attributes can be collected leading to incomplete data or missing information not transferred through the collected attributes.[86]

The redlining example shows a second problem. Attributes, or a group of attributes, can be a proxy for protected grounds, which are not allowed to be used in decision-making. The area codes used in redlining were also strongly correlated with certain ethnic groups, as the ethnic composition of neighbourhoods often reflects segregation. Redlining had also strong discriminatory effects towards these ethnic groups, and is therefore considered illegal. This means that to avoid discrimination it is often not enough to take out features containing protected grounds, like ethnic background or sex. Correlated proxies, like postcode, can take over their role and make discrimination possible. To prevent discrimination also these proxy attributes have to be taken out or their discriminatory role has to be tackled in another way.[87]

Subjective bias can also be introduced in the data through the labelling of the data. In this case the data collection method is not biased, but the labels, or outcomes for the target variable, itself represent biased decisions. When the labels reflect objective measurements or observations, this will not be an issue. Even when the data is collected in a biased way, it will represent true observations. E.g. a technical test for alcohol intoxication will still be representing an objective result, even when racial profiling was involved in the data collection. But labelling involving the subjective assessment of humans can also translate prejudices of these humans into the data. E.g. historical data on hiring or admission decisions will reflect prejudices involved in these procedures. The distinction between both sources of bias is important as it will need a different approach to correct it. When certain populations are over- or under-represented, this can be corrected through introducing

[85] Barocas, Solon and Selbst, Andrew D., "Big Data's Disparate Impact", op. cit.
[86] Calders, Toon, Indrė Žliobaitė, “Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures”, op. cit.
[87] Barocas, Solon and Selbst, Andrew D., "Big Data's Disparate Impact", op. cit.

weighting factors to correct the distribution, while biases reflected in the labelling are less easy to determine and correct.[88]

Lastly, these range of sources concerned unintentional discrimination. But all these factors can also be consciously manipulated to mask discriminatory intentions.[89] The legal prohibition of indirect discrimination is targeted both at such masked intentional discrimination as against unintentional discrimination.

This overview shows that data mining can in several ways deliver discriminatory results. When these results are used in decision-making concerning persons, they can lead to discriminatory practices.

3.4.4 Technical measures and anti-discrimination by design

The potential for discriminatory results of big data practices can be addressed through technical measures. The attention for discrimination in the data mining literature is very recent. Technical measures and design approaches to prevent discrimination are less developed than concerning privacy, but are partly inspired by the privacy-by-design developments. Technical measures have been developed for 2 distinct tasks: to discover discrimination and to prevent discrimination. Here we present the state of the art.

Discrimination discovery aims at discovering in historical decision records discriminatory practices or contexts in which discrimination takes place. This is important as these archived decision data can be used to train data mining algorithms in decision support systems, which would then exhibit the same discriminatory biases present in this data. This is not an obvious or easy task. These decision records can be high-dimensional and by consequence represent a large amount of contexts. It is possible that discrimination only takes place in some very specific contexts which are hidden in a broad range of situations. Checking some specific situations can be done with generally used statistical methods, but discovering niches of discrimination is still an unsupported task. Further, it concerns mostly indirect discrimination, or apparently neutral situations or practices which result in discrimination, which means again that we are confronted with a wide range of potential discriminatory situations.

The approach taken in the 'Discrimination Discovery in Databases' or DCUBE-system[90] extracts classification rules from the historical data and filters these rules based on risk measures. The classification rules are extracted through association rule mining, enabling to extract all rules of a certain form which are supported in the dataset by a minimum amount of decisions. These rules are filtered by looking at the change of confidence level when a potential discriminatory term is removed from the decision rule. When the confidence level

---

[88] Calders, Toon, Indrė Žliobaitė, "Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures", op. cit.

[89] Barocas, Solon and Selbst, Andrew D., "Big Data's Disparate Impact", op. cit.

[90] Ruggieri, S., D. Pedreschi, F. Turini, "DCUBE: Discrimination Discovery in Databases", ACM International Conference on Management of Data (SIGMOD 2010): 1127-1130. *ACM*, June 2010; Pedreschi, D., S. Ruggieri, F. Turini, "The Discovery of Discrimination", in Custers, Bart, Calders, Toon, Schermer, Bart, & Zarsky, Tal (Eds.), op. cit., 2012.

of the rule rises with the addition of a discriminatory term, the rule is considered to be potentially discriminatory and the data supporting it to be representing potentially discriminating decisions. This measure concerns direct discrimination, similar measures have been developed for indirect discrimination.

Another approach uses the idea of situation testing. For each member of a potentially discriminated group which obtained a negative outcome, other individuals with similar characteristics are searched in the decision set. If the decisions for the individuals which also belong to the same group are significantly more negative compared to those for the other individuals which are not part of that group, the decision can be considered potentially discriminating. This method is dependent on the ability to discern the members of a specific group in the dataset.[91]

Several methods are also developed to prevent data mining algorithms from discriminating. Aim is to have discrimination-free classifiers, or to learn a model from data which labels new data as accurately as possible, but where these labels are also uncorrelated with protected grounds. Accuracy is an obvious objective, but blind optimisation would lead to reproducing the discriminations present in the training data. The extra objective is in trade-off with accuracy, as these techniques prevent discrimination but with the cost of lesser accuracy. An obvious approach would be to remove attributes reflecting sensitive criteria or which are highly correlated with them. But such approach can also remove too much useful information and have a strong effect on the accuracy of the models. Some more advanced techniques have been developed, which can be categorised depending on the phase in which they intervene in the data mining process:
1. Pre-processing techniques remove or compensate the discrimination from the training dataset. This adapted training set can then be used with standard data mining methods and the resulting model is assumed to be discrimination-free. Several techniques to remove discrimination from the data set have been developed, of which we give some examples. 'Massaging' selectively relabels some of the instances in the dataset. In other words, some negative results of the discriminated group are changed in positive and some positive of the other group are changed in negative, while keeping the overall distribution of results intact. By selecting those instances closed to the decision boundary, that is those instances which would flip their result with the smallest change, the amount of changes and the impact on accuracy is minimized. 'Reweighing' leaves the input data intact, but changes the impact of data by assigning different weights to records. Positive results in the protected group get a higher weight than negative results, while negative results in the other group get a higher weight than positive results. Not all data mining algorithms can work with weighted data and an alternative is to first construct a new dataset by resampling the original dataset using this weighting technique. The weights are now reflected in the probabilities used in such resampling. 'Resampling' leads to a duplication of some records of positive results with the protected group and negative results of the other group, while dropping some of the other results. In the adapted dataset the distribution of results is changed to eliminate discrimination.

[91] Luong, B. T., S. Ruggieri, F. Turini. "k-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention", 17th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2011): 502-510. *ACM*, August 2011.

2. In-processing techniques which change the learning procedures themselves by restricting their search space to models which are non-discriminating. Such restricting is done by embedding measures of discrimination aside of the original measures into the data mining algorithms. This adds a penalty for certain areas of the search space, which are optimal for the viewpoint of the original accuracy measure, but which represent discriminatory practices. Optimisation in the data mining algorithm is now not only done for the original accuracy measure but also for the discrimination measure. In the literature we found several adaptations of standard data mining techniques based on this principle.[92]
3. Post-processing techniques which adjust discriminatory models. Here models are made using training data without adaptation and standard data mining techniques, which therefore will exhibit the discriminatory practices present in the data. But these models get extra corrections to compensate or remove this discrimination. Kamiran et al. present a model correction method of decision trees. Decision trees partition the space of instances in non-overlapping regions, and which get labelled according to the majority label. Some of these regions can be very homogeneous with respect to sensitive attributes and represent potential discrimination. The correction method consist in searching which regions can be relabelled with the least loss in accuracy and the highest decrease of discrimination. Such post-processing methods for discrimination seem to be rarer in the literature till now, but model correction methods have been explored in several other settings.[93]

A limitation present in these methods is that they treat all differences in treatment as unjustified and discrimination which has to be removed. This while it is possible that some of the differences can be objectively explained from other, justified factors. E.g. a difference in income levels between certain groups can point to discrimination, but also differences in education level. Such differences in education level can point in itself to discrimination in the education system, but an employer who bases his hiring decisions on education levels is not illegally discriminating even when the end result leads to a much more limited presence of the more vulnerable group among his employees. In this sense a certain amount of discrimination can be justified, as follows from the definition of indirect discrimination. This legal definition entails the possibility that a particular disadvantage, resulting from an apparent neutral practice, is "objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary". When there is a part of the difference in treatment which can be justified, then removing all difference from the data results in training a model which exhibits positive discrimination. In other words, what is to be considered indirect discrimination present in the data and what are justified differences in treatment, is an important aspect for preliminary consideration before choosing which algorithm to apply. Conditional non-discrimination methods aim at measuring the explainable and non-explainable parts of the discrimination and at building classifiers which only remove the non-explainable, and therefore unjustified, part. This is done by splitting the data in groups conditioned on the variable which explains the justified difference in

[92] Kamiran, Faisal, Toon Calders, Mykola Pechenizkiy, "Techniques for Discrimination-Free Predictive Models"; Sara Hajian, Josep Domingo-Ferrer, "Direct and Indirect Discrimination Prevention Methods" in Custers, Bart, Calders, Toon, Schermer, Bart, & Zarsky, Tal (Eds.), op. cit., 2012.
[93] Kamiran, Faisal, Toon Calders, Mykola Pechenizkiy, "Techniques for Discrimination-Free Predictive Models", op. cit., 2012

treatment. Next pre-processing techniques, like massaging or resampling, are applied locally within every group.[94]

Kamiran et al. compared the methods presented here. The pre-processing proved able to significantly reduce the discrimination with minimal effects to accuracy, while the in-processing methods were not very effective. The post-processing methods could reduce the discrimination even further, while accuracy got slightly more affected but remained on an acceptable level.[95] These results show that effective prevention of discrimination in data mining is possible, even while this field of research is still in its infancy stage.

An approach similar to privacy-by-design, which will be presented in the next chapter, is not yet developed to deal with anti-discrimination. What is available now is a set of algorithms which make it possible to check for discrimination and to prevent discrimination. Further elements, like established technical objectives, design patterns, design methodologies, auditing procedures or assessment methodologies are not developed yet. This resembles the early stage of privacy-by-design, when only a range of PETs were available without integrated design approaches. An important field of work therefore remains. However, this work does not need to start from scratch. Just like privacy-by-design methodologies are an extension of methodologies used for security, a similar extension can be developed for anti-discrimination. Most effective is to broaden and further develop the privacy-by-design approach and its methodologies to include anti-discrimination objectives. Again, this cannot be limited to a technical approach, but needs to be taken up in a broader framework for transparency and accountability.

3.4.5 Adapting the legal and institutional framework concerning anti-discrimination

Making big data free of discrimination is more than a technical question. First, the equality bodies will have to take up the task to drive the development and promotion of anti-discrimination-by-design as part of their mainstreaming efforts. Further, they will have to develop a transparency and accountability framework which will be partly based on anti-discrimination legislation but also on the data protection framework. For this they will have to cooperate with data protection authorities. Lastly, they will have to engage in a prospective societal and political debate on the 'new discriminations' made possible by big data. This will have to lead to the development of new norms, be they social, professional or legal.

EU anti-discrimination law is quite different from the data protection framework. Both frameworks deal with the protection of fundamental rights but the anti-discrimination framework is more narrow in scope compared to the data protection framework. In terms of protection the anti-discrimination framework is focussed on post facto responding to

[94] Žliobaitė, Indrė, Faisal Kamiran, Toon Calders, "Handling Conditional Discrimination", Data Mining (ICDM), 2011 IEEE 11th International Conference on. IEEE, 2011, pp. 992-1001; Kamiran, Faisal, Indrė Žliobaitė, "Explainable and Non-Explainable Discrimination in Classification" in Custers, Bart, Calders, Toon, Schermer, Bart, & Zarsky, Tal (Eds.), op. cit., 2012.

[95] Kamiran, Faisal, Toon Calders, Mykola Pechenizkiy, "Techniques for Discrimination-Free Predictive Models", op. cit., 2012.

discrimination. It does provide legal redress against discrimination, of which, as mentioned earlier, the area of application varies according to the protected ground. Aside of this reactive role, which was deemed not effective enough, a second, preventive approach was developed. Mainstreaming stands for including equality objectives (in practice focussed on equal treatment of women and men) into all general policies and not restricting anti-discrimination policy to specific areas or to providing legal redress. This entails a shift from combating discriminatory practices to a focus on the outcome of policies. To put this into practice equality bodies play an important role. Such equality bodies are only provided in the directives concerning sex and race, although their task can be enlarged at the national level. Neither do these directives provide a detailed description of tasks and competences, which led to a much diversified landscape of equality bodies among the member states. Compared to data protection authorities (DPA) the equality bodies have a more limited role. Generally they do not have the competences to control and investigate practices companies or other actors like the DPAs have. Equality bodies only work reactively in getting legal redress or support efforts by individuals to do so. Both provide guidance but also here we see differences. DPAs mainly give advice on how to apply data protection law. Similarly equality bodies engage in networks with stakeholders to promote best practices in applying the directives. But equality bodies have in their mainstreaming role also the function to do research and give policy recommendations, and some provide awareness-raising activities and recommendations outside the area of application of the anti-discrimination directives.[96]

Notwithstanding these differences both the equality bodies and the DPAs can take up the task to address discrimination through data mining. DPAs have no specific anti-discrimination role but they control if processing of personal data is done for legitimate purposes and according to the data protection principles. Further, the processing of sensitive data, which include all the protected grounds apart from sex, is forbidden with some exceptions. This gives the DPAs the competence to address discrimination due to the processing of personal data. Equality bodies on the other hand can take up this issue as part of their mainstreaming efforts. To our knowledge no substantial work has been done on big data and data mining by the equality bodies. Cooperation between DPAs and equality bodies to share expertise and build a common agenda seems necessary to give anti-discrimination in data mining and big data adequate attention.

In order to prevent discrimination through data mining and big data it will be important to develop an adequate transparency and accountability framework. In the context of administrative decision-making Citron warns strongly for how procedural safeguards are imperilled by automation in what she calls the automated administrative state. Involving more humans in the decision-making process will not do, Citron states, since in practice the difference between automated decision-making and non-automated decision-making with human intervention before the decision is very little. People, including state officials, are often fooled by technology or do not possess the capacity or information to do a proper assessment of the computer-generated suggestion. Too much trust in the results provided by the computer leads to an 'automation bias'. Such exaggerated trust in computers runs

[96] Ammer, Margit et al., *Study on Equality Bodies set up under Directives 2000/43/EC, 2004/113/EC and 2006/54/EC*, Human European Consultancy & Ludwig Boltzmann Institut für Menschenrechte, October 2010.

counter to the duty of care. The EUCJ case C-503/03 between the European Commission and Spain, concerning the refusal of entry into the Schengen area based on flagging in the SIS system, confirms this duty of care. The EUCJ pointed out that such refusal without preliminary verification if the person presented an actual danger violated EU law. Further, coding can result in a hidden form of rule-making through mistakes or interpretations by programmers. The construction of a profile can be seen as a similar form of hidden rule making. Citron shows that procedural safeguards need a major overhaul to remain effective. She points out the need to give several of these safeguards a particular implementation. First, the opacity of automated processes imperils the information to the subject and the proper review of elements on which the decision (in the case of profiling to flag a person as a risk) is based. Therefore proper audit trails should be required as part of the system, documenting the rules applied and data considered. Decision-makers should also explain in their motivations in detail how they relied on computer-generated information. Decision-makers should also be trained in how to evaluate and scrutinize such information, in order to combat automation bias. Further, automated systems should be designed with transparency and accountability as inherent objectives. Citron advises to make code available as open source. In the case of profiling that would imply making the algorithms and profiles open for review. As such code or algorithms can function as hidden rule-makers, this provides an alternative way for external scrutiny. For the same reason, she advises to allow the public to participate in reviewing and commenting on the systems. Lastly, she makes a plea for proper testing.[97] A similar plea can be made for more transparency and accountability towards the use of decision-making based on or assisted by data mining and big data in the private sector.

The legal grounds for transparency are not present in the anti-discrimination but in the data protection legislation. As will be presented in the next chapter, these transparency obligations mostly concern the actual personal data used and are less clear and stringent concerning information on how the data is used. Nevertheless, the DPAs have the competence to further develop these obligations into clear requirements towards data controllers. The WP 29 made clear that further processing of personal data with the purpose to make inferences concerning the individuals to inform measures or decisions with regard to these individuals needs informed consent by these individuals. And such informed consent implies that data subjects need access to their profiles and to the logic used to develop such profile. A full access to the algorithm is not required. Although we fully agree with the plea of Citron, these algorithms will often be considered as confidential commercial information. But a clear account of the decisional criteria needs to be given.[98] The new General Data Protection Regulation (GDPR) will probably enlarge the requirements for transparency concerning profiling. In this context the development of standardised auditing tools can be a method to provide adequate transparency without needing access to the algorithm itself.

The current data protection directive gives data subjects the right to object to automated decision-making which produces legal effects concerning him or significantly affects him,

[97] Citron, D.K., "Technological due process", *Washington University Law Review*, 85, 2008, pp. 1249-1313.
[98] WP 29, Opinion 03/2013 on purpose limitation, WP 203, 2 April 2013, pp. 45-47.

when this decision is solely based on automated processing to evaluate personal aspects relating to him. In a range of cases this right to object does not apply when adequate measures are present to safeguard the individual's legitimate interests, like the ability for the individual to present his view.[99] This rather weak protection will be strengthened in the GDPR, although how remains under discussion. But in all cases such protection is dependent on assuring transparency on the decisional criteria and safeguards against unintended discrimination. The DPAs can assure a better implementation of the transparency obligations of data controllers and enforce that also decisional criteria and precise information on how personal data is used is given to data subjects. In cooperation with the equality bodies they can take the initiative to ensure that adequate auditing tools are developed and used, and that privacy-by-design methodologies and data protection impact assessments include attention to anti-discrimination.

To conclude we recommend that the equality bodies take up the task to drive the development and promotion of anti-discrimination-by-design as part of their mainstreaming efforts. Further, they will have to develop a transparency and accountability framework which will be partly based on anti- discrimination legislation but also on the data protection framework. For this they will have to cooperate with data protection authorities. Cooperation between DPAs and equality bodies to share expertise and build a common agenda seems necessary to give anti-discrimination in data mining and big data adequate attention.

3.4.6 Preparing to deal with new discriminations

Lastly, the equality bodies will have to engage in a prospective societal and political debate on the 'new discriminations' made possible by big data. In particular the healthcare case study made clear that big data leads to new 'visibilities' which also makes new discriminations possible. Genetic information and other health-related information can reveal a lot of information on future health situations. This makes personalised healthcare possible, but also more personalised risk calculations which can be used in insurance, employment decisions and so on. Will an employer still hire a person or provide him with the needed training if genetic data reveals that this persons has more chance to develop a disabling or life-threatening disease in a couple of years? Is an insurance company allowed to take certain health information into account to differentiate risk premiums? Further, does the availability of such information strengthen the personal responsibility to take this into account and adapt his lifestyle? Can a person be considered liable if he does not and therefore being excluded from certain insurances? All these cases concern new forms of rational discrimination becoming possible due to big data. This time the grounds for such rational discrimination are not any more the protected grounds like sex and ethnic origin, but a range of genetic predispositions, possibly combined with lifestyle choices. This requires new norms concerning the grounds on which rational discrimination is permitted and when it is not. Where to draw the lines of what is acceptable and what not needs to be subject of a wider societal and political debate. Equality bodies have a lot of expertise on

[99] 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 October 1995, article 15.

where such questions do appear and can therefore play an important role to organise and drive such debate.

Conclusions

The benefits of big data are similar in the economic and the social domain. Big data can lead to improved efficiency and innovation. The best practices identified in the part on economic externalities are therefore also applicable to capture benefits in the social domain. Analysing how big data affects interaction between actors provides a better understanding of how big data produces these benefits, as well as how it creates negative externalities. Big data makes it possible to deal with a higher amount of interactions, such that transaction costs get lowered while retaining the granularity needed for more individualised and targeted responses. This lowering of transaction costs to process and analyse a large amount of data also makes producing fine-grained results possible. This points to interoperability as a key enabler of big data practices. Best practices concern therefore the improvement of interoperability in its several dimensions: semantic, technical, legal and organisational.

However, big data also creates negative externalities as a consequence of how big data changes interactions and of the inadequacy of legal frameworks to deal with these changes. The improved awareness provided by big data, and the decision-making resulting from it, can be used for positive and negative purposes and therefore lead to discrimination. Further, when data becomes a network good, old discriminative practices reflected in the data can proliferate and lead to unintended discrimination.

Data mining can result in unintended discrimination due to choices in the problem definition and to discriminatory bias in the training data. Effective methods to deal with such bias have recently been developed and are object of further research. These methods need to be further developed and integrated in auditing tools and in an anti-discrimination-by-design approach, similar as in privacy-by-design.

The equality bodies will have to take up the task to drive the development and promotion of anti-discrimination-by-design as part of their mainstreaming efforts. Further, they will have to develop a transparency and accountability framework which will be partly based on anti-discrimination legislation but also on the data protection framework. For this they will have to cooperate with data protection authorities. Cooperation between DPAs and equality bodies to share expertise and build a common agenda seems necessary to give anti-discrimination in data mining and big data adequate attention.

Big data provides a larger and more fine-grained visibility of persons on a range of issues. This enlarges the capacity to make distinctions between people, which can be very intrusive and give rise to new forms of differentiated treatment or new discriminations. New norms on what are acceptable grounds for decision-making will need to be developed. Therefore, the equality bodies will have to engage in a prospective societal and political debate on the 'new discriminations' made possible by big data.

4 LEGAL EXTERNALITIES

4.1 INTRODUCTION

In the case studies two main legal frameworks were put forward as presenting legal externalities of big data. First, intellectual property rights were seen as an important barrier to big data practices in most of the case studies (crisis informatics, energy, environment, healthcare). It restricts access to data and its mechanisms to obtain authorisation for the use of the data do not scale well and function badly in a big data context. The question can be raised if the balancing done in the current IPR framework still functions and if the framework needs updating to become better adapted to a data economy. Second, the data protection framework was also considered outdated. Although its objectives are widely shared, its current mechanisms also pose application problems in the context of big data. A third framework, the protection of trade secrets, will also be considered. This did not figure as a legal externality in the case studies, but it is a potential best practice to address the trust problems raised in a B2B context. Uncertainty and fear of manipulation resulting from information leakage could also be encountered in a purely business context like the energy case study or in the mixed citizens and business context of smart cities. A legal protection of trade secrets can restore the gatekeeping capacity of organisations.

A 'legal externality' is not a current and well-defined concept. For the purposes of this report it is defined as negative impacts of big data on the functioning of a legal framework and negative impacts of this legal framework on big data operations. These frameworks could also be considered under economic, social and ethical externalities. The choice to consider them as legal externalities is due to the fact that a lot of the issues raised have to do with the actual functioning of the legal framework. For each legal framework the current framework will be evaluated as well as possible solutions to make it function better in a big data context.

4.2 INTELLECTUAL PROPERTY RIGHTS

4.2.1 Introduction: intellectual property rights on data

Intellectual property rights (IPR) protect intellectual creations and reserve certain exclusive rights concerning their use and distribution to their creators or those to whom these rights have been transferred. Each regime defines what falls under its protection. Certain regimes can apply to data and datasets. Most relevant are copyright and database protection.

Copyright can exist over the individual data as well as over the database as a whole. Copyright protection for databases results from the copyright for collections. It is provided by directive 96/9/EC of 11 March 1996 on the legal protection of databases and protects databases where "the selection or arrangement of their contents" is a result of "the author's own intellectual creation"[100]. The protection concerns the organisation and structuring of

[100] European Parliament and the Council, Directive 96/9/EC of 11 March 1996 on the legal protection of databases, art. 3.1

the data but does not extend to the individual data items itself. Copyright of individual data items grants exclusive rights on the individual item, but is independent from the copyright over the database structure as a collection. Both have to be checked separately and can belong to different rights holders. General principle of copyright is that it protects expressions, but not ideas in itself, nor procedures, methods or mathematical concepts.[101] Aim is to protect products of human intellect and creativity. Trigger for the protection is therefore some sort of originality. Originality implies originating from an author, but also being the result of some intellectual or creative effort.[102] Novelty is not required, but the mere investment of effort in copying information does not reach the threshold for copyright protection. Also, purely factual information is not protected under copyright, seen that facts are discovered and not the result of creativity. Copyright protection given to the expression does not extend to the underlying facts. This factual information can be used by others, as long as they do not reproduce it in the protected expression. The application of copyright on datasets is therefore not straightforward: not all data is protected by copyright, but only those that meet the originality-requirement.

Directive 96/9/EC also provides another form of protection for databases as a sui generis right. This sui generis-right protects the maker of a database who has made a substantial investment in the creation of a database. Only the costs associated with "obtaining, verification or presentation of the contents" as a whole are taken into account, not the cost associated with obtaining, creating or updating individual data items.[103] No originality is required, protection is based on the investment. The maker of the database is given the right to prevent extraction and re-utilization of the whole or of a substantial part of the contents of that database. This right does not prevent lawful use, consisting of extracting or re-utilizing insubstantial parts of database contents. In other words, a query over a database is as such no violation of the sui-generis database protection and does not need a specific authorisation by the right holder, except when the response contains the whole or a substantial part of the contents. The substantiality can be assessed both quantitatively and qualitatively. Further, this use may not conflict with the normal exploitation of the database or unreasonably prejudice the legitimate interests of the maker of the database. Result is that any big data processing involving a substantial part of a database will need permission from the right holder during the 15 years term of protection. The sui-generis protection and the protection by copyright can coincide.

Copyright and the sui generis-protection of databases balance different interests in order to get the best result on the macro-level. Copyright is meant to provide an incentive for innovation by assuring the remuneration of the creative person for his effort. Therefore a temporary monopoly right is granted to an author or creator. This exclusivity permits an author to gain back his costs, as otherwise others could just compete with him based on the mere cost of copying his work. No protection would mean that the price would be set in a competitive market at the marginal cost of producing an extra copy and make it impossible to recuperate the earlier investments. Such a situation would present a strong barrier

[101] WIPO Copyright Treaty, art. 2; TRIPS, art. 9 §2.

[102] Goldstein, Paul, *International Copyright. Principles, Law, and Practice*, Oxford University Press, New York, 2001, p. 161-164.

[103] Truyens, Maarten & Patrick Van Eecke, "Legal aspects of text mining", *Computer law & security review* 30 (2014), 160.

against innovation. Copyright drives the price up, but furthers innovation and the end result should be advantageous for society at large and present a positive externality. This is only true when the protection is not too strong and becomes itself a barrier for innovation, by limiting the availability and use of the protected work for new innovations too strongly. A too strong or too wide protection can therefore also stifle innovation and lead to negative externalities. The question which amount of protection provides the optimal balance from a societal point of view is not easy to answer.

4.2.2 Evaluation of the current framework

The EU introduced the sui generis protection of databases based on this assumption that protection by property rights was needed to attract investments.[104] In the US the Supreme Court refused in its Feist-decision such protection based on investments or 'sweat of the brow'. It was motivated by a similar economic reasoning but considered such protection as a barrier to the exchange of ideas and information. Copyright protects only expressions and not facts, and this freedom to keep the underlying information available for immediate use was considered essential. Result is that databases are more strongly protected in the EU than in the US, and this protection presents a stronger barrier to the interoperability of data and the utilisation of datasets in big data practices.

The question is who struck the balance right and when does protection leads to negative externalities? In this section the available evidence will be evaluated. The next section will look into potential solutions based on this evaluation.

The available evidence does seem to point out that the legal protection of data and databases is too strong in the EU and leads to negative externalities. In its own evaluation in 2005 the Commission could not find a positive economic impact of the directive on database protection. After an initial rise in production of databases this amount had in 2004 fallen back to pre-directive levels. However, the empirical data was doubted to be fully representative and uncertainty about the actual evolution remained. Notwithstanding this doubt, the same data showed that database production in the US, where less legal protection exists, had continuously grown in the same period.[105] Although this conclusion is not drawn in this evaluation report, in our opinion this gives at least an indication that stronger protection of databases does not attract investments and even can have a negative effect. The European Commission has kept the directive unchanged seen the large support of the concerned industry for the directive. Ian Hargreaves qualified this in his review of the intellectual property framework as an example of policy development inconsistent with the available evidence.[106]

A similar conclusion can be drawn from the evaluation of the economic impact of the recent Spanish 'Snippet'-law, which does not concern database rights but ancillary copyright. It included in the Spanish Law on Intellectual Property an article stipulating that news

[104] European Parliament and Council, Directive 96/9/EC of 11 March 1996 on the legal protection of databases, recital 12.
[105] 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. 18-20.
[106] Hargreaves, Ian, *Digital Opportunity: Review of Intellectual Property and Growth*, May 2011, p.19.

aggregators on the Internet (e.g. Google News) did not require an authorisation for making non-significant snippets of news and other informative articles available to the public, but that such use is subject to the payment of a fee to the right holder. This right to this fee is irrevocable. It is therefore not possible for authors or internet publications to waive this rights. The motivation behind this law is that news aggregators are rent-seeking on local news publications. A substitution effect is presumed where some customers are satisfied with the snippets and do not click through to the original source. Through this substitution effect news publications would attract less public and news aggregators would catch a part of the advertisement revenues which would otherwise go to the publishers. News aggregators would cause negative externalities at the publishers. The fee is considered to compensate for this negative externality. Opponents of this law hold that news aggregators have no substitution effect, but instead a market expansion effect. These services make it easier to find news items and reduce search times. As they enable readers to consume more news, news aggregators would not lead to less but rather more traffic to the news publications. Overall result would be one of market expansion and of positive externalities for the publishers. Which theory is correct, can be derived from empirical observation of the traffic. A recent study looked into the effect of this law and noticed a significant negative impact on Internet traffic to news publications. This law targeted in the first place Google News and in reaction it has withdrawn from Spain, as it did before when similar laws were introduced in other countries. On average traffic to news publications went down with 6% in the 3 months after this law entered into force on 1 January 2015. Smaller, lesser-known publication have been more affected with traffic losses up to 14%. In other words, the total market seems to shrink and also gets more concentrated. Less traffic will lead to less advertising revenue, which can in the longer term threaten the existence of some of these publications. Further, not just Google News has withdrawn but also several local news aggregators have closed down.[107] Although these results reflect only a short period, they are significant enough to draw some conclusions. The rent-seeking argument that Google News and other news aggregators attract public and accordingly advertising revenues which would otherwise go to the news publications does not seem to hold. The fact that the traffic shrinks after the disappearance of news aggregators rather points to the market expansion effect of these news aggregators. These empirical observations also concur with 8 other studies on the effect of news aggregators referenced in this report. The law further stifled innovation, as it led to the closure of several news aggregators, including some very innovative ones. We can therefore conclude that the stricter copyright law does not strike the right balance between the opposing interests and negatively impacts on all market players. These results reflect the positive network effects of data. News aggregators have a positive result on the overall market as they create new data from other sources and make those other sources more usable and therefore valuable.

All these results seem to point to the negative effect of overprotection of data sources through intellectual property rights. Access to data sources is too much restricted and this blocks innovative uses of these sources.

[107] NERA Economic Consulting, *Impacto del Nuevo Artículo 32.2 de la Ley de Propiedad Intelectual*, 9 July 2015.

4.2.3 Adapting IPR to big data

The BYTE case studies showed that IPR presented significant barriers for big data operations. Also the literature study in D2.1 and policy review in D1.2 uncovered specific problems for data mining. One solution is to change the legal framework and limit the reach of the protection by copyright and database rights. Solutions within the current legal framework are to make datasets more accessible through collective licensing and open licenses. The structured stakeholder dialogue Licences for Europe, held in 2013, addressed several issues concerning copyright, like user-generated content (UGC) or data- and text mining (limited to text and data mining for scientific research purposes). But exactly these issues proved the most difficult.[108] We present the different solutions.

Solution 1: collective licensing

Collective licensing tries to find a solution for the transaction cost problem present in the copyright framework, while continuing to assure revenues to the creator. In fact, the transaction cost problem has been present within copyright already before the advent of digitalisation and big data. Conceptually copyright remains based on the transaction-model, implying a contract between two parties after a negotiation. But this model became quite soon in its history more an assumption than a reality. Reason is that copyright expanded from a property right over a creation, which could change owners in one transaction, into a bundle of rights, which includes a range of derivative or ancillary rights over discrete types of use made of that creation. Examples of these derivative rights are performing in public, communicating to a public through broadcasting, making copies, renting works, etc. Historically this bundle of rights developed through adaptation to new technical forms of creation and dissemination. Copyright fragmented in a bundle of rights and involved an augmented, continuously ongoing set of transactions. In other words, copyright became more similar to dealing with the delivery of a service. This fragmentation of copyright made in practice individual negotiations and transactions impossible and entailed high transaction costs. Rights clearance became much more complex and collective management of rights became the logical solution. Collective rights management organisations (CMOs) perform the rights clearance for a range of creative works and rights. Transactions are not negotiated any more but become adhesion contracts. In the literature this shift to collective management has also been designated as a shift from a property or transaction model towards a liability model. A liability model implies that the use of the work creates the right on a fee for the copyright holder, but in practice he does not consent to a contract any more with the user of his work.[109]

The adaptation to digitalisation reinforced this shift away from the transaction-model even more. The service character became even stronger by the dropping of the right of exhaustion for digital copies. Second-hand books can be sold without a need for an

[108] European Commission, On content in the Digital Single Market, COM(2012)789, 18.12.2012; Results can be found on http://ec.europa.eu/licences-for-europe-dialogue/en

[109] Gervais, Daniel, "The Changing Role of Copyright Collectives" in Daniel Gervais (ed.), *Collective management of copyright and related rights*. Kluwer Law International, 2010, pp. 3-36; Schovsbo, Jens. "The Necessity to collectivize copyright-and dangers thereof", available at SSRN 1632753 (2010).

authorisation of or a fee to the author. In the digital world it is more difficult to differentiate between reproducing and transferring a copy, and therefore between an original and a second-hand e-book. The right to make a private copy has come under pressure for similar reasons. Digitalisation also led to a second important change. Earlier rights clearance did not occur so much with the general public, but with other companies or professional actors which then communicated the work to the general public. In the digital age rights clearance and enforcement of copyright is targeted directly to the general public, which perceives this as a continuous interference with its use. In other words, the actual rights clearance mechanisms worked in a B2B environment, but are not very well adapted to a mass customer environment. The struggle of the content industry and the collective rights management organisations against illegal downloading and copying is well known and it would lead us too far to discuss it. But some lessons for the area of data can be learned from the observation that a large part of the problem originated from the fact that users were often not able to access the copyrighted material in a lawful way or to obtain the right to use it beyond the primary use of viewing, reading or listening. In other words, to the lack of realistic licensing options. This results from the fragmentation of rights over specific uses, often managed by different CMOs, and over jurisdictions with their own specific CMOs. Further CMOs or other industry actors like publishers often held onto old business models and therefore blocked new uses. However, the justification of CMOs relies on their role as effective brokers, which make works available while limiting transaction costs. This analysis is mostly based on more traditional content markets, like music and movies, but it is relevant valid for the market in data.[110]

CMOs have not had an important role in the data market before, but licensing plays a large role in access to scientific publications and commercial databases. In this case publishers play the role of CMO by providing access to a range of publications or to specific databases. Complaints have been raised about the too expensive or too limited access, and over the availability for new uses like data mining. Scientific publishers have the role of organising large-scale availability and access to scientific information, but their performance in the new digital environment can be questioned in a similar way as for CMOs. In the stakeholder dialogue Licences for Europe several scientific publishers promised to introduce a “licensing clause for subscription-based material”, enabling researchers to mine for non-commercial scientific research purposes as part of the subscriptions done by the universities, together with “a web-based 'mining portal'”.[111] Licensing also plays an important role in social media like Twitter and Facebook and other Internet services like Google maps or YouTube, as we illustrated in D1.2. E.g. users of Twitter grant a “worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute”.[112] Twitter makes these tweets available on its own terms to others, regulating that access through technical means (its API) and legal means (its Terms of Use). In other words, social media companies and Internet services play a similar role as a CMO in organising access to content on a large scale.

[110] Ibid.

[111] European commission, Licences for Europe. Ten pledges to bring more content online, 13 November 2013, http://ec.europa.eu/internal_market/copyright/docs/licences-for-europe/131113_ten-pledges_en.pdf

[112] Twitter, Twitter Terms of Service, version May 18, 2015, https://twitter.com/tos; Lammerant, Hans et al., *D1.2 Big Data Policies*, BYTE project, 31 August 2014, 75-83.

In order to make licensing more efficient, new licensing models have been developed to ensure availability and access. An important example is the extended repertoire system or extended collective licensing (ECL), which was originally developed in the Nordic countries. CMOs have reacted to the difficulties in advance clearance of rights with repertoire licenses, which allow the user to use any work in the repertory of works held by the CMO. Extended collective licenses, when made possible in the copyright law, go a step further in the collective management of rights. These licenses combine a voluntary assignment of rights by rights holders to the CMO with an extension of the repertoire to non-member rights holders. When a substantial number of rights holders join forces in or with a CMO, the law extends the repertoire of this CMO automatically to all rights holders in the same category. These other rights holders can still opt out of this extended collective licensing scheme. But participation in the transaction becomes an 'opt out'-choice instead an 'opt in'.[113] Using such ECL in the context of data would come down to the 'opt out'-possibility (through a robots.txt-file) as provided by Google and which it used in its defence in the court cases concerning Google News.

Collective management of rights can go even further and shift completely into a liability model with no involvement of the original creator. One option is a compulsory licence. In fact, the recent Spanish Snippet law comes down to such a compulsory licence, as it provides that the fee cannot be waived. In this case a form of opting out is technically still possible through the use of a robots.text-file, but the freedom of choice by the creator is thereby limited to the choice between making its work unavailable for news aggregators and receiving a fee according to the law. An even further-going liability mechanism is putting a fee on or taxing the blank media through which dissemination can take place, like audio cassettes in the past and hard drives and memory sticks nowadays. These levy systems are widespread and meant to compensate for private copies.[114] In the EU schemes vary a lot and are not harmonised, although the ongoing copyright review aims to do so. Looking to the context of data, an option could be to use a similar levy system on processing power, as this would correlate most with the capacity to use data sources on a large scale. Still, such a system would raise important practical questions, e.g. which are the rights holders (as not all data falls under copyright).

Another way to compensate for the obstacles presented by copyright are open licenses. Open licenses remain within the copyright framework, but use it for the opposite purpose: to allow that data can be freely used. I will not deal with this extensively as it has been discussed earlier in D1.2. The development of these licences can surely be considered as a strong contribution and a best practice in the current context. But it is a best practice which needs some caveats as given by Creative Commons, one of the main developers of open content licenses: "CC licenses are a patch, not a fix, for the problems of the copyright system. … Our experience has reinforced our belief that to ensure the maximum benefits to both culture and the economy in this digital age, the scope and shape of copyright law need to be reviewed. However well-crafted a public licensing model may be, it can never fully achieve what a change in the law would do, which means that law reform remains a

[113] Gervais, Daniel, "The Changing Role of Copyright Collectives", op. cit., 2010, pp. 3-36.
[114] For an overview: WIPO, International Survey on Private Copying, Law and Practice 2013.

pressing topic. The public would benefit from more extensive rights to use the full body of human culture and knowledge for the public benefit. CC licenses are not a substitute for users' rights, and CC supports ongoing efforts to reform copyright law to strengthen users' rights and expand the public domain".[115]

Although the presented collective licensing schemes would guarantee a remuneration for the data producers and curators while reducing the transaction costs involved, the question remains if such solutions strike the best balance on macro-level in terms of promoting innovation and are therefore 'best practice'. The earlier mentioned evaluation of the Spanish Snippet law was not very promising in that respect. These solutions do not take the market expansion due to positive network effects into account. Further, the volume of data involved in data mining and big data can make big data practices very costly once its use has to be remunerated. Remuneration for these uses can therefore stifle big data practices.

Solution 2: limiting the extent of IPR

This points to the second option: reforming the copyright framework in order to limit the extent of exclusive rights over data and enlarge the free access to it.

On the legislative level several options exist to open up the copyright framework. They imply providing access to data for specific big data and data mining purposes without remuneration. A first possibility would be to withdraw the sui generis database protection and in that way align the protection of databases with the US and other regions. This would make the protection for databases more limited. Protection would exist only for those databases which can be considered original and therefore make more databases available for free use. However, it does not solve the problems for data mining over the databases which are still protected under copyright. At the moment this option of abolishing the sui generis database protection has not been raised at the political level.

Another option is to create more exceptions to the reserved uses in the copyright and the database protection framework, which would create more space for data uses without a need for an authorisation or a license. Especially for data mining such an exception would be useful. Several countries did notice the problem the IPR regime poses for text and data mining and adapted the exceptions or are discussing a change. The UK added an exception for computational analysis for the purpose of non-commercial research. In its international strategy on IPR the UK government further included the aim "to secure further flexibilities at EU level that enable greater adaptability to new technologies".[116] Japan updated its copyright law with a new exception giving space for information analysis.[117] The Australian Law Reform Commission recommended after a public consultation to adopt a general 'fair

[115] Creative Commons, "Creative Commons and Copyright Reform", 16 October 2013. http://creativecommons.org/about/reform

[116] Intellectual Property Office, *The UK's International Strategy for Intellectual Property*, 11 August 2011, p. 13.

[117] 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, pp. 10-11.

use'-exception like in the US as a flexible and technology-neutral solution.[118] A similar review took place in Ireland, leading to the recommendation to add exceptions for 'content-mining' for purposes of education, research or private study to both copyright and the protection of databases as well as a fair use-exception.[119] Both recent reports have not led yet to legislative action. Two main options seem on the table. Adding a fair use exception, similar like in the US, would create the space to have copyright more adaptive to new technological changes. Data mining could be considered as such a fair use. However, more adaptive does not mean less conflict, as also in the US many of such new uses became the subject of disputes in the courts. The other option is to add a specific exception for text and data mining. This has the advantage to be clear, but it is perhaps also too hesitant and limited, in particular as the current proposals limit the exception to non-commercial uses. The European Commission plans to adapt the exceptions regime for text and data mining in its upcoming proposals to review the copyright framework.[120] Compared with the licensing solutions, these legal changes would lead to a limitation of the possibilities to ask for revenue for the creation of data. On the other hand, the overall effect could be one of market expansion due to the larger use made of the data.

Conclusions

We can conclude that intellectual property rights, licensing and contracts are important tools in regulating access and use of information. Evaluation of the current copyright and database protection shows that this framework is too restrictive. Possible solutions involving legal change are to drop the sui generis-right on databases and to add new exceptions for data and text mining in copyright law. A solution within the current legal framework is collective licensing, which lowers transaction costs but preserves remuneration. Extended collective licenses can make this framework even more flexible. However, the evaluations show that limiting copyright is preferable, as it leads to market expansion due to network effects.

4.3 THE PROTECTION OF TRADE SECRETS

Big data practices lead to much more interactions and visibility to the outside world. Individuals experience this as a privacy problem, but it is also a problem for companies. Leakage of confidential information to competitors can severely harm a company, while also leakage to commercial partners can create problems. The fear for such leakage and its possible consequence of harm or loss of autonomy can lead to a reluctance to engage in business processes with commercial partners which involve large data flows. In this situation big data practices are only taken up when they can be fully or to a large extent internalised and data leakage to the outside world or other companies can be avoided or remains exceptional. In these cases the uncertainty leads to an external effect of less investment in big data practices. When such leakage or the consequences can be contained

[118] Australian Law Reform Commission, *Copyright and the Digital Economy. Final Report*, ALRC Report 122, 30 November 2013, p. 13.
[119] Copyright Review Committee, *Modernising Copyright. The Report of the Copyright Review Committee for the Department of Jobs, Enterprise and Innovation*, Dublin, 2013.
[120] European Commission, A Digital Single Market Strategy for Europe, COM(2015)192, 6 May 2015.

this would create an incentive to engage in big data-based business processes with other commercial partners as well and lead to more economic efficiency or innovation at the overall societal level. This 'privacy' problem of companies can be addressed by the protection of trade secrets. This legal protection erects new legal barriers to information outflow. It makes it possible to fine-tune both the access to data and the permitted use of data. This restores the capacity for effective gatekeeping by organisations and enables them to participate in practices with larger data flows or interactions between partners.

4.3.1 A new legal framework to protect trade secrets

The European legal landscape concerning protection of trade secrets and confidential information is very diverse. The European Commission has proposed recently a directive introducing protection of trade secrets, which is now under consideration in the European Parliament.[121] Trade secrets protection is not an intellectual property right as such, but it fulfils a similar and complementary role to IPR in protecting intellectual assets. Important difference is that it gives no exclusive right on the use of the information. But it gives protection against appropriation with dishonest means. It can therefore also be characterised as a protection against unfair trade practices.[122] The diversity of national protections for trade secrets results from this mixed character as it developed from different legal frameworks, IPR or unfair trade practices, in these jurisdictions.

A trade secret is defined in the newly proposed directive as information which is secret in the sense that it is not generally known or readily accessible, which has commercial value because of being secret, and which has been subject to reasonable steps to keep it secret. It can concern a wide range of information, from technical information like production processes to commercial information like client lists, as long as its secret character provides it a commercial value. It can therefore also concern all sorts of data flows which a company wants to keep confidential. Protection is only given against unlawful acquisition, use and disclosure and as long as it is kept secret. When the holder of the trade secret does not take adequate measures to ensure its confidentiality, he loses the protection. In case of unlawful acquisition, use or disclosure the directive provides civil law remedies for redress, including interim and precautionary measures.

Acquisition is unlawful when it is done by unauthorised access or copy of the documents or materials which contain the secret or from which it can be deduced, through a breach to a confidentiality agreement or another duty to maintain secrecy, through conduct contrary to honest commercial practices or through theft, bribery or deception. Use or disclosure of a trade secret is unlawful when done by a person which has acquired the trade secret unlawfully, who is in breach of a confidentiality agreement or duty to secrecy or who is in breach of a duty, contractual or other, to limit the use of the trade secret. The use or

[121] European Commission, Proposal for a Directive of the European Parliament and of the Council on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure, COM(2013) 813, 28.11.2013

[122] Reichman, Jerome H., "How trade secrecy law generates a natural semicommons of innovative know-how", in Rochelle C. Dreyfuss, Katherine J. Strandburg (eds.)-*The Law and Theory of Trade Secrecy. A Handbook of Contemporary Research*, Edward Elgar Publishing Limited, Cheltenham, 2011, pp. 185-200.

disclosure of a trade secret remains unlawful when it has been obtained from a third person who unlawfully used or disclosed it. This is a very wide definition of unlawful behaviour, but it is limited by excluding a range of lawful acquisitions, uses and disclosures. E.g. the acquisition is legal when done by independent discovery or creation or by observation, study or reverse engineering of a product. In other words, no exclusivity for the use of the trade secret is provided. When others acquire the trade secret themselves in a lawful manner, which includes reverse engineering the product of others, then the trade secret loses its protection. Further the protection of the trade secret is denied when the disclosure is part of a legitimate exercise of the freedom of expression and information, for revealing misconduct or illegal activity, for protecting a legitimate interest, and some other circumstances. However, this proposal has been criticised for providing a too broad protection and stifling legitimate disclosure, like by journalists, NGOs or whistle-blowers and this will surely remain a hot discussion point in the treatment by the European Parliament.

For our subject the importance of this proposal is that it provides a legal framework to restore boundaries by defining a range of unfair trade practices concerning information. This legal framework can reduce some of the uncertainty associated with the larger openness to commercial partners through big data practices. Important is that the reasonable steps to keep information secret do not have to be purely technical or physical and have to be considered in the specific circumstances. Allowing access to information or giving visibility to internal processes, but regulated through a contract which strictly limits for which purpose this information can be used or contains confidentiality clauses, can be a reasonable substitute to technical protections which also lock every external partner out. In other words, this legal framework gives some teeth to non-disclosure agreements and other agreements concerning information. This is especially important because of the shift from exchanges of goods limited in time to continuous services involving permanent data streams from sensors or user data. The momentary character of transactions of goods provided a natural protection which now gets lost. The service-orientation requires new organisational and contractual set-ups, for which the new legal framework on trade secrets provides a part of the legal infrastructure. Therefore new legal protection of trade secrets figured also in the policy recommendation of the German working group on Industrie 4.0.[123]

We can conclude that some version of protection of trade secrets is a necessity in the context of big data. It restores the capacity for effective gatekeeping by organisations and enables them to participate in practices with larger data flows or interactions between partners.

4.3.2 A toolbox for fine-tuning data flows

A full protective mechanism to ensure control over and protection of information for companies cannot be limited to legal frameworks as such. It needs to be further developed in legal tools, like contractual arrangements, and organisational and technical measures, which can also be standardized. At the moment this toolbox is still underdeveloped for big data, although it is possible to build upon work done in related areas like IT security, clouds

[123] Forschungsunion and acatech, *Umsetzungsempfehlungen für das Zukunftsprojekt Industrie 4.0*, http://www.bmbf.de/pubRD/Umsetzungsempfehlungen_Industrie4_0.pdf

or privacy. The technical toolbox developed as part of privacy-by-design (e.g. privacy-aware data mining) or for IT security in general (e.g. encryption, access control measures) can be adapted in this context as well.

Further development of contractual set-ups can be done through drafting standard contractual clauses, like concerning confidentiality of data and its use in big data applications. Such standardized contractual arrangements are in particular needed in cloud environments when several software layers provided by different actors are involved. A lot of work has already been done on standard service level agreements (SLA) for clouds. In December 2012, the European Commission (EC) and the European Telecommunications Standards Institute (ETSI) launched the Cloud Standardization Coordination (CSC) initiative. A first phase resulted in an end report defining roles in cloud computing and categorising more than 100 use cases. It also inventoried standardisation work and identified specific needs. This report concluded that relatively few standards exist concerning SLAs for cloud services and that the main requirement for standardisation in relation to Service Level Agreements was the creation of an agreed set of terminology and definitions for Service Level Objectives, with an associated set of Metrics for each service level objective. This would make it easier to create SLAs with a clear and comparable description of cloud services.[124] This requirement was addressed in the SLA Standardisation Guidelines, resulting from the work of the Cloud Select Industry Group – Subgroup on Service Level Agreement (C-SIG-SLA). These Guidelines contain a useful overview and categorisation of service level objectives, which need to be included in SLAs and which can be further developed. It contains 4 major categories of service level objectives: Performance, Security, Data Management and Personal Data. Specifications of purposes and rights to use data can be found in the Data Management category under data classification, with further specifications of the data life-cycle under other sub-topics. Personal data receives a more specific treatment.[125] This work is now continued in a second phase of CSC work, of which the draft reports are now available. No specific work is done on SLAs, but one of the reports concerns further defining requirements concerning interoperability and security and this includes how to address it in SLAs.[126]

This work produces very valuable high level definitions and requirements, on which technical, organisational and legal measures can be further developed. Similar work would be very useful in other areas of big data processing and it can be seen as a best practice to address in a coherent way underdeveloped areas of concern. A lot of attention goes to security and to dealing with personal data. No specific attention is given to confidentiality of non-personal data. A lot of the concerns about confidentiality of commercial relevant data can probably be addressed by the work done on security and privacy, but specific attention to the legal aspects would probably be valuable. The interplay between legal, organisational and technical measures will become clearer in the next part on privacy. The overall framework for the protection of personal data is much better developed. Further

[124] ETSI, "Cloud Standards Coordination, Final Report", version 1.0, November 2013

[125] C-SIG-SLA, "Cloud Service Level Agreement Standardisation Guidelines", Brussels, 24 June 2014

[126] ETSI, "Interoperability and Security in Cloud Computing, Analysis, conclusions and recommendations from Cloud Standards Coordination Phase 2"

developing such a framework to address the confidentiality of non-personal but valuable data can be an important tool to foster the adoption of big data.

Conclusions

The case studies showed that companies can have 'privacy' problems as well. Some legal means for gatekeeping are needed to create legal certainty and overcome distrust and reluctance. The protection of trade secrets can be a useful tool to create legal certainty in the shift to services. It makes fine-tuning possible of both the access to data and the permitted use of data. This restores the capacity for effective gatekeeping by organisations and enables them to participate in practices with larger data flows or interactions between partners. Such legal instruments have to be further developed in standardised contractual arrangements, as part of a toolbox of legal, organisational and technical safeguards.

4.4 PRIVACY AND PROTECTION OF PERSONAL DATA

Privacy concerns were raised in all case studies, except the energy case study which concerned a purely B2B context. At the same time the data protection framework was perceived as an inefficient and outdated framework. Similar criticisms have been raised in the scholarly literature and by industry. These critics present data protection as an obstacle for big data practices and the scientific and economic advantages it can bring.[127] Others consider data protection broken and not effective any more to protect privacy in the age of big data.[128]

First we will consider the question on how to apply data protection in a big data context. We will consider the criticism on the European data protection framework and the scientific debate on anonymisation which has generated a lot of this criticism. Critics have promoted a risk-based approach, in which the application of data protection shifts from the collection to the use of the data and which is modulated on the risks involved. The WP 29 has countered this view by showing that the current data protection principles can be implemented in a risk-based approach without limiting the extent of the protection. In such a risk-based application of data protection law it is important to evaluate the total set-up of technical, legal and organisational safeguards, and not just look at the theoretical risks associated with anonymisation techniques or other technical safeguards. Legal tests, like concerning identifiability, have to assess the actual implementation, taking into account all safeguards, instead of assessing the technical impossibility of identification. What has to be

[127] 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); Rubinstein, Ira S., *Big Data: The End of Privacy or a New Beginning?*, NYU School of Law, Public Law Research Paper No. 12-56; Moerel, Lokke, "Big Data Protection: How to Make the Draft EU Regulation on Data Protection Future Proof", Tilburg University, 2014.

[128] 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 Daniel J. Solove, "The PII Problem: Privacy and a New Concept of Personally Identifiable Information", *New York University Law Review*, December 2011, 1814-1894; Mantelero, Alessandro, "The future of consumer data protection in the EU Re-thinking the "notice and consent" paradigm in the new era of predictive analytics.", *Computer Law & Security Review* 30.6 (2014): 643-660.

assessed is the actual risk in the particular set-up and not the theoretical risk associated with the techniques involved.

Secondly, privacy-by-design is an important approach to develop practical solutions for data protection in big data settings. To make an overall evaluation possible of the risks involved, it is important to develop standard solutions with known capabilities and risks for the whole range of technical, legal and organisational safeguards. The objectives of the data protection framework have to be translated into design objectives and integrated in design methodologies and development processes. Such design methodologies should be broadened from purely technical to include also legal and organisation safeguards. Privacy-by-design moves a large amount of evaluation and decision-making from the individual acts of data processing to an aggregate assessment in the design phase. Standardisation makes swift risk analysis and auditing of systems possible, thereby lowering transaction costs involved.

Thirdly, the current data protection framework still contains individual transaction-based elements which are difficult to scale, e.g. consent. Where possible, these need to be substituted by more aggregated and collective mechanisms. This also entails a shift from an individual control model to a liability model with a more prominent role for regulators.

4.4.1 The application of data protection law to big data

First we will present some views on the difficulties in the application of data protection law on big data. These turn around the notions of personal data and anonymisation, purpose limitation and data minimisation. Generally proposals are put forward for a risk-based approach to address these problems. But the notion of 'risk-based approach' turns out to cover very different proposals. Next we present the position of the WP 29, which has given guidance on how these data protection principles and notions can be applied, including in a context of big data.

Industry views

Industry voices have stated that data protection law is outdated and not adapted to an environment where data is much more shared. Data protection law was developed when personal data after its collection remained mostly in data 'silos' of governments and companies and its use could be more easily determined at the moment of collection. Nowadays data resources can be more easily connected and re-utilised, and therefore new uses of personal data can pop up in a much later stage. Typical examples are big data practices, where data resources are 'explored' for its value for new uses. These practices get blocked by a data protection law that is too restrictive when they were not foreseen at the moment of collection. Therefore, data protection principles like purpose limitation and data minimisation are seen as too restrictive. Instead, these industry voices have pleaded for a more risk-based approach, where risks are assessed at the moment of use instead of the moment of collection. This makes a more contextual assessment possible of the level of protection needed. Further, data protection law has been presented as too binary. Data is personal or not, triggering for the application of data protection law as one-size-fits-all solution. In opposition to such extensive application of a strong regime of protection

whenever data concerning an identifiable person is used, a more contextual and gradual protection, based on the possible harm, is put forward.[129] In other words, this risk-based approach entails modulating both the field of application and the level of protection of personal data according to risk.

The anonymisation debate

An important question is if the notions on which data protection law is based are still applicable in the context of big data. Data protection law uses the notion of personal data as a trigger for the application of data protection, and anonymisation as a tool to withdraw data from this legal regime. Once anonymisation becomes impossible, all data linked to an identifiable person would in principle remain under data protection law. Research showed that standard anonymisation methods were much less reliable than assumed and could not guarantee full anonymisation, in particular when a lot of data is available. Within the computer science and the statistical community this has led to a contentious debate on how to deal with these results. This debate got reflected in the legal community and legal scholars started to discuss if data protection law could remain in its current shape. Legal scholars like Paul Ohm came to the conclusion that privacy law was broken, seen the fact that fundamental notions of data protection law like personal data or personally identifiable information are based on the distinction with non-personal data and the possibility to anonymise personal data. The EU data protection framework is applicable on "any information relating to an identified or identifiable natural person".[130] Recital 26 of the directive 95/46 makes clear the link with anonymisation, by stating that "the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable".[131] Ohm points out that, when reidentification becomes easy, the range of what is personal data, and therefore also the area of application of data protection law, can potentially become boundless.[132] He and other scholars started to discuss versions of risk-based approaches, where protection was modulated according to harm or risk.[133] Here we will first give an introduction into the technical problem of anonymisation. Next we will present some scholarly views on how to deal with it. All promote a risk-based approach, but what such approach entails differs a lot.

Anonymity and anonymisation techniques

[129] World Economic Forum, prepared in collaboration with BCG, "Unlocking the Value of Personal Data: From Collection to Usage", World Economic Forum, February 2013; DIGITALEUROPE, "DIGITALEUROPE Comments on the Risk-Based Approach", 28 Augustus 2013.
[130] 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 October 1995, article 2(a).
[131] Ibid., recital 26.
[132] Ohm, Paul, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization", op. cit., pp. 1735-1745.
[133] Ohm, Paul, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization", op. cit.; Schwartz, Paul M. and Daniel J. Solove, "The PII Problem: Privacy and a New Concept of Personally Identifiable Information", op. cit.

Anonymity is not limited to being unknown, in the sense that a person's name is not known. When someone can be tracked through his behaviour, e.g. on the Internet through cookies or on CCTV images through biometric identification, a person's name can remain unknown but he can be singled out, monitored and actions can be targeted to this person, be it targeted advertising or security checks in public areas. The WP 29 has clarified that "'identification' not only means the possibility of retrieving a person's name and/or address, but also includes potential identifiability by singling out, linkability and inference." The WP 29 defines singling out as "the possibility to isolate some or all the records which identify an individual in the dataset". Linkability is the possibility to link 2 or more records concerning a data subject or a group in the same of different datasets. E.g. through the presence of unique combinations of non-unique attributes, or so-called quasi-identifiers, in distinct datasets or records. Inference is "the possibility to deduce, with significant probability, the value of an attribute from the set of other attributes".[134] Anonymisation techniques have to be evaluated according to these three aspects.

A basic step to anonymise data is to remove all direct identifiers, like names and identification numbers. In other words those elements that directly point to a unique individual. But this necessary first step has proven to be not enough. Some attributes are not unique, like birthday, postcode or gender, but their combination can be unique. These attributes become quasi-identifiers and make it possible to single out an individual. They can be used to link information present in distinct datasets when both datasets contain these quasi-identifiers. If the extra dataset also contains the direct identifiers, it is possible to reconnect the anonymised set with these direct identifiers. In other words, even after the removal of directly identifying information, there can be 'auxiliary information' available which can be identifying in combination and lead to privacy breaches. A famous example is the reidentification of Governor Weld in deidentified hospitalisation records with the use of publicly available voter registration records by Latanya Sweeney. These and similar attacks on deidentified data showed the problem that anonymisation is not perfect and can be broken in some cases. On the other hand, this does not mean that all data in such a dataset can be reidentified.

A range of methods have been developed to deal with this 'auxiliary information'-problem. Two broad categories of techniques can be discerned, methods based on randomisation and methods based on generalisation. Randomisation changes the data in such a way that the link between the data subjects and the data is broken, while leaving the macro-level characteristics intact. Techniques include: noise addition or modifying attributes by adding some noise while assuring that the general distribution is retained, permutation or the shuffling of attribute values (such shuffling retains the general distribution of values since no value is changed), differential privacy or a query based form of noise addition (properties of the whole set can be queried, while the privacy of the individuals in the set is preserved). Differential privacy is a criterion which ensures that the probability of any element to belong to a certain set, like the results of a query, differs not significantly from the probability of not belonging to it. Noise is added with each query to assure this criterion is satisfied. To guarantee privacy in a provable manner queries may not be considered independent and earlier query results have to be taken into account. Generalisation aggregates individuals in

[134] WP 29, Opinion 05/2014 on Anonymisation Techniques, WP216, 10 April 2014, p. 10-12.

the dataset by modifying the scale or order of magnitude of the attribute values of the quasi-identifiers. E.g. using the year of birth instead of the date of birth makes larger aggregations of individuals and makes it more difficult to single them out. A range of measures have been developed to guide such generalisation. K-anonymity groups an individual with at least k-1 other individuals. Main problem with k-anonymity is that it does not prevent inference attacks and sensitive information can be inferred when such a sensitive attribute has the same value or at least with a high probability in some groupings. Additional measures try to prevent the occurrence of groupings with low attribute variability.[135] The WP 29 concluded that all these anonymisation techniques, as well as pseudonymisation, failed to fulfil its three evaluation criteria. All methods had their advantages but also their shortcomings or limitations.

Risk-based approaches

Not everyone shares the pessimistic view put forward by Paul Ohm on the impact on data protection of this research about reidentification. The Canadian Information and Privacy Commissioner Ann Cavoukian made a strong rebuttal of the pessimistic views towards anonymisation and showed that anonymisation is not as broken as feared. She points out that breaches of anonymity and reidentification only succeeded in a limited amount of cases. Even in the cases published, most of the data remained anonymous. In practice the likelihood of reidentification is small for most data and it are mostly the outliers which are at risk for reidentification.[136] El Emam et. al. made a systematic review of the reidentification literature and came to the conclusion that most of these studies concerned small-scale studies on data that was not deidentified according to current standards. This is the case for several famous early studies, like the reidentification of Governor Weld in deidentified hospitalisation records with the use of publicly available voter registration records by Latanya Sweeney.[137] The risk for reidentification would have been greatly reduced through better generalization. These studies brought the problem under attention, but were followed by a lot of research on improved anonymisation methods. Perfect anonymisation was not achieved, but clear progress was made and anonymisation techniques greatly reduce the risk for re-identification in practice.

However, not all criticism can be so easily dismissed. Especially for high-dimensional, sparse datasets anonymisation proves difficult and the existing techniques are inadequate. This is also admitted by Cavoukian and Castro.[138] Sparse datasets are sets with a high amount of attributes or dimensions, while individuals all have small amount of non-zero values compared to the amount of dimensions. E.g. social network structures where everyone only has a limited amount of connections or interactions with people compared to the total amount of people, shopping data where everyone only buys a limited set of goods compared to the total amount of goods on offer. In such datasets it is very difficult to

[135] WP 29, Opinion 05/2014 on Anonymisation Techniques, WP216, 10 April 2014, p. 12-19.

[136] Cavoukian, Ann, Daniel Castro, *Big Data and Innovation, Setting the Record Straight: De-identification Does Work*, Information and Privacy Commissioner, Ontario, Canada, 16 June 2014

[137] El Emam, Khaled, Elizabeth Jonker, Luk Arbuckle, Bradley Malin, "A Systematic Review of Re-Identification Attacks on Health Data", *PloS one* 6.12 (2011): e28071.

[138] Cavoukian, Ann, Daniel Castro, *Big Data and Innovation, Setting the Record Straight: De-identification Does Work*, op. cit., p. 3.

discern a limited number of quasi-identifiers or to make a distinction between quasi-identifiers and sensitive information. Any attribute can be identifying in combination with others and is therefore a quasi-identifier, while it often also can avail sensitive information.[139] This is also made clear by the research of Yves-Alexandre de Montjoie who shows that only a few data points in a set of mobility data or of credit card transactions are enough to single out an individual.[140] The standard anonymisation methods have been developed to anonymise tabular datasets with a relative limited amount of dimensions, but they do not work well for these high-dimensional datasets. And when big data techniques are utilised to process personal data, it concerns often such high-dimensional forms of data.

For this reason computer scientists like Arvind Narayanan advise to take a more principled and scientific approach to privacy. The pragmatic, risk-oriented approach is denounced as similar to the penetrate-and-patch approach which proved ineffective in the field of computer security. . Instead Narayanan et al. advise to learn from cryptography research and to develop more rigorous concepts of provable privacy.[141] They argue for a precautionary approach. The precautionary principle comes, just like 'risk-based approaches', in many shades. It ranges from very weak, implying that regulation is permitted when risks are uncertain, to much stronger versions, where an actor has to prove that the risks are acceptable before being allowed to act. Precautionary approaches shift the burden of proof, in this context it means that data controllers carry the burden of proof concerning the risks involved in the processing. Narayanan et. al. argue that current standards in the US are too lenient. Now a data processor releasing data can shift the burden of proof when he uses standard anonymisation techniques, but seen the shortcomings signalled above Narayanan et. al. want to move the standard upwards and put a higher burden of proof on the data controllers. They agree that current ad-hoc deidentification is a useful safeguard but not sufficient. A precautionary approach using provable privacy methods is needed, in particular for releases of data. With this plea Narayanan et. al. adopt EU-like standards of non-identifiability and go even further than the European reasonable effort-test in their demand for scientific rigor. Narayanan et. al. prefer to use this stricter burden-of-proof standard as a guiding principle to develop solutions more tailored to the specific dataset and the circumstances. A one-size-fits-all legal requirement they deem to lead to sub-optimal results. More tailored implies also more narrow. Unlimited releases to the general public are considered to be too risky in general. Instead they advise to use legal agreements which restrict the use and further flow of data to other parties. Also they advise to increase transparency by requiring in privacy policies also the disclosure of transfers of anonymised data, and not just personal data or PII. One example of provable privacy in a context of more narrow access they present is differential

[139] Narayanan, Arvind, and Vitaly Shmatikov. "Myths and fallacies of personally identifiable information." *Communications of the ACM* 53.6 (2010): 24-26; Rubinstein, Ira and Woodrow Hartzog, "Anonymization and Risk" (August 17, 2015). *Washington Law Review*, Vol. 91, No. 2, 2016, pp. 9-11.

[140] de Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). "Unique in the Crowd: The privacy bounds of human mobility." Scientific reports, 3; de Montjoye, Y. A., Radaelli, L., & Singh, V. K. (2015). "Unique in the shopping mall: On the reidentifiability of credit card metadata.", *Science*, 347(6221), 536-539.

[141] Narayanan, Arvind, Joanna Huey, Edward Felten, "A Precautionary Approach to Big Data Privacy", 19 March 2015, 8-9.

privacy. As mentioned earlier, this is a query-based method of noise addition. Important is that it is based on a well-defined metric of privacy and that is takes into account the amount of information released through subsequent queries. General access to the dataset through a release of the data is not given, but the actual access to the data through queries is monitored to continuously assess the privacy risk.[142]

Ira Rubinstein and Woodrow Hartzog also present a risk-based approach, but one which is more risk-tolerant.[143] In contrast with Narayanan they want a lower standard which shifts attention from output to process. Rubinstein and Hartzog advocate to focus on the process of minimising risk, instead of preventing harm. They draw the conclusion from the anonymisation debate that focussing on the perfect outcome of anonymisation does not lead to useful regulatory approaches. Changing the focus to the process of risk management would lead to a more flexible and workable legal framework for data releases. Harm-based regulation does not function well, as harm for badly deidentified data and data releases is often difficult to detect and to quantify. On the other hand, other approaches like focussing on transparency or on consent would also not bring positive results, as they put all responsibility on individuals with a limited capacity to control all the information. Instead Rubinstein and Hartzog recommend to focus on processes needed to limit the risks on reidentification and sensitive disclosures. It is easier to ensure that data controllers follow sound procedures that limit the risk, than to ensure specific outcomes like preventing reidentification. A process-based regime gives custodian-like responsibilities. Liability flows from disrespect of standard procedures and of these responsibilities, even when no harm occurred. For this approach they draw inspiration from data security law in the US, which is characterized by regular risk assessment, the implementation of technical and procedural safeguards and on adequate response in case a system is compromised. Very specific legal rules are avoided, instead reasonable adherence to industry standards is the norm and criterion to determine liability. Such a reasonableness standard allows to be more sensitive to contextual factors and leaves room for fast adaptation. On the other hand, such an approach is more risk-tolerant and acknowledges that perfection is not possible. Rubinstein and Hartzog see a similar evolution in how the FTC, based on consumer protection law, deals with data releases. The FTC considers data not 'reasonably linkable', and therefore not subject to additional data protection measures, when a company takes reasonable measures to deidentify the data, publicly commits not to reidentify the data, and contractually prohibits recipients of the data to attempt reidentification. Tene and Polonetsky make a similar plea for a more risk-based and scaled approach to personal data and point to this FTC practice, which shifts the inquiry from a factual test of identifiability (to be understood as a purely technical assessment) to a "legal examination of an organization's intent and commitment to prevent re-identification".[144] Rubinstein and Hartzog do not want to go as far as the pragmatic approach of Cavoukian and Castro, in the sense that they recognize the weaknesses of anonymisation. But in their view the debate has been so far too much focussed on anonymisation and they recommend to use the full spectrum of safeguards in tailoring a company's obligations to the risks involved. Data controls through

[142] Narayanan, Arvind, Joanna Huey, Edward Felten, op. cit., 2015.
[143] Rubinstein, Ira and Woodrow Hartzog, "Anonymization and Risk", op. cit.
[144] Tene, Omer, and Jules Polonetsky, "Big data for all: Privacy and user control in the age of analytics.", op. cit.

contracts and access control mechanisms are considered as important as deidentification. They advise, just like Narayanan et. al., to limit or eliminate the release-and-forget model for data releases. In particular they point to other access regimes which can be used to modulate the risks, like direct access and query-based access. Direct access works with authorizing data access with a data use agreement or license. These agreements can include security measures, prohibitions to redistribute or to reidentify data and limit the use to specific purposes (e.g. research). Direct access methods allow any query and give full access to the results, but limit the risk for linkage attacks by not releasing data sets to the general public nor for general use. Query-based access allows to run queries but in a controlled environment. Several methods are possible. One method grants access to confidential data, but limits both the queries which can be posed and the access to the results. Access to the underlying data is not given. Another approach is differential privacy, where query results are altered through noise addition and taking into account the data already released over subsequent queries. Also in this case no direct access to the underlying data is given. As an example Rubinstein and Hartzog refer to the controlled access used in several cases of genetic research, which include the use of deidentification, tiered access depending on the sensitivity of the data, and informed consent. Rubinstein and Hartzog highlight from these cases that it is possible to capture most benefits from open access without releasing the data to the general public or without restrictions.

These views show there are many risk-based approaches possible from very strict, asking for provable privacy, over pragmatic versions to very open, where data collection is free but risk is assessed when the data is used.

WP 29 and the risk-based approach

The WP 29 has formulated a range of guidelines on how to apply data protection to big data. According to the WP 29, a risk-based approach is already part of the current data protection framework. But it is clear that the WP 29 has a different understanding of what such a risk-based approach entails than its critics. In its Statement on the role of a risk-based approach in data protection legal frameworks the WP 29 makes clear that a risk-based approach concerns compliance mechanisms, but not general data protection principles and data subject rights. The WP 29 takes position against the plea to shift the focus of regulation from collection of personal data to its use. Such plea comes down to modulating not just the compliance mechanisms according to the risk, but limiting its field of application.[145]

In its rebuttal of this plea, the WP 29 points to the fact that protection of personal data is a fundamental right and that therefore any processing should respect this right (and remain to be subjected to the data protection framework). According to the WP 29 none of the data subject's rights can be modulated according to risk and implicitly the same is true for the applicability of the data protection framework in itself. What can be modulated according to risk are the compliance mechanisms and accountability obligations. Similarly the fundamental data protection principles are applicable whatever the processing and the risk, but their implementation can take the nature of the processing and the risk into account.

[145] WP 29, Statement on the role of a risk-based approach in data protection legal frameworks, WP 218, 30 May 2014.

E.g. documentation of processing can vary according to the risk, but some documentation is necessary to provide transparency and accountability.[146]

In its guidance the WP 29 makes clear how data protection law can be applied in a risk-based approach. First we will look at how the WP 29 deals with the anonymisation debate. Next we will look at how the WP 29 deals with the other contested notion of purpose limitation.

WP 29 on anonymisation

The WP 29 also adheres to a pragmatic view on anonymisation. It uses a strict standard in applying the data protection principles, but is pragmatic in its evaluation of anonymisation techniques. Full anonymity, where the data protection framework does not apply any more, entails that the person, which is concerned in the data, cannot be identified any more. The WP 29 made clear this means not just that the data cannot be linked to a name or other direct identifier, but also that it is not more possible to single out the individual, to link records relating to that individual or to infer information concerning that individual. Such deidentification has to be irreversible.

Recital 26 makes clear the determination of (un)identifiability does not need to be as certain or rigorous as a mathematical proof, but has to take into account "all the means likely reasonably to be used" in a possible identification attempt. This refers to potential deidentification efforts by the controller or third parties. This reasonable effort-test refers to the concrete efforts needed for potential reidentification, e.g. in terms of cost, know-how and computing power needed, likelihood of identification seen the availability of other data. Therefore the WP 29 does not require an absolute and provable impossibility of identification but it looks to the practical effort needed for a successful attempt of reidentification. Raising that effort can be done not just by anonymisation but also by a broader range of technical and organisational safeguards. In this way the limitations of anonymisation techniques do not make the definition of personal data boundless. It is the actual effort needed for reidentification in the context of all the safeguards in place, and therefore the actual risk for reidentification, which defines what is personal data and therefore also which processing is subjected to data protection law. More precise criteria of privacy and anonymity are very useful to guide these assessments, but provable privacy is not an absolute requirement in this approach.

In case full anonymity is not possible, the data protection rules keep applying on the data. In this case anonymisation techniques remain useful as a technical safeguard. The assessment of the anonymisation efforts becomes part of the compatibility assessment of further processing, which we will consider in the next section. The criteria to judge identifiability (singling out, linkability and inference) become criteria to judge the overall effectivity of technical and organisational safeguards in that assessment. In this sense a technique like pseudonymisation clearly is not an adequate anonymisation technique, but it can be a useful technical safeguard to minimise the access to the data and to prevent certain uses alongside other safeguards and within the application of the data protection framework.

[146] Ibid.

These factors determining the effort needed for reidentification can change over time and therefore also the determination of the identifiability, which implies that this assessment has to be regularly repeated.[147] Therefore the WP 29 advises against a "release and forget"-approach. The risk for identification has to be monitored and assessed regularly.[148] The availability of other data can make reidentification easier and can therefore have an impact on this assessment. When a third party receives an anonymised dataset, it is allowed to process it without applying data protection requirements provided that it cannot identify the data subjects in the dataset. But this third party has to take all contextual factors, like the shortcomings of the anonymisation technique used, into account when deciding how to use the data. E.g. when combining the dataset with other datasets in its possession which would create a risk for identification, the processing falls again under the application of data protection law.[149] In other words, when a context of big data enlarges the risk for reidentification a stronger set of safeguards, be it anonymisation techniques or other, will be needed. In this approach the failure of anonymisation techniques to guarantee anonymity does not make the application of data protection law impossible or boundless. It rather requires a broader set of safeguards, depending on the risk involved.

These safeguards also impact the assessment of which further processing is found compatible with the purposes of the data collection and modulate this compatibility according to risk. The criteria developed by the WP 29 for anonymisation techniques are also useful to assess the whole setup of technical and organisational safeguards, in which the use of anonymisation is one safeguard among others.

WP 29 and purpose limitation

In its Opinion on purpose limitation the WP 29 clarified the principle of purpose limitation and gave guidelines on its application on big data. It pointed out that this principle already contains the means to modulate its implementation according to risks, but the principle as such remains applicable in all circumstances. The principle consists of 2 main building blocks:
1. purpose specification: personal data may only be collected for 'specified, explicit and legitimate' purposes
2. further use or processing is only allowed when not incompatible with those purposes.
The purpose of the collection of personal data has to be defined before or at the moment of the collection, but not later. The WP 29 makes clear that purpose specification is an essential first step to apply other data protection safeguards. Without a sufficient specification of the purpose it is not possible to make proportionality assessments, that is to determine which collection and further processing of personal data is necessary, relevant and adequate. The specification of the purpose makes the processing of personal data predictable and provides legal certainty. Other safeguards like transparency and user rights are only fully effective when a purpose is specified and thereby enables data subjects to

[147] WP 29, Opinion 05/2014 on Anonymisation Techniques, WP216, 10 April 2014, p. 8-10.
[148] Ibid., p. 24.
[149] Ibid., p. 10.

check compliance and to determine their own interest vis-a-vis the processing.[150] Purpose specification is therefore not a principle which can be dropped, while keeping the other data protection principles. Also more diluted implementations, like allowing very broad purposes, affect the implementation of other data protection principles.

Important is that further use for other purposes is not forbidden as such, but only when the purposes of this further use are incompatible with the original purposes. Directive 95/46 does not impose a requirement of the same or a similar purpose, but rather prohibits incompatible purposes. The second building block limits, but also provides some room for processing with altered or new purposes. The notion 'further processing' consist of any processing after the initial processing act of collecting the data, at which moment also the original purpose is defined. Therefore the compatibility of this original purpose and the purpose of the further processing acts has to be assessed. In some cases this can lead to the conclusion that even a change of purpose is permissible.

The WP 29 has provided extensive guidance on how to make a compatibility assessment. It recommends not to limit such an assessment to a formal assessment comparing the usually written statements, but to make a substantive assessment beyond the formal statements, taking into account the context and other factors. Inherent in such a multi-factor assessment is that negative aspects can be compensated by better performance on other aspects or extra safeguards. Key factors to consider in the compatibility assessment are:
a) The relation between the purpose of the original data collection and the purpose of the further processing. Again this should not be limited to a textual comparison, but focus on the substance of the relation between the purposes. Is the further processing already implied in the original purposes, a logical next step, or is there a more distant, partial or even non-existent link? In general, the more distance between the purposes, the more problematic for the compatibility assessment.
b) The context of the data collection and the reasonable expectations of the data subjects concerning its further use. Main question is how would a reasonable person expect its personal data to be used in the same context of the collection and of the situation of the data subject? Important aspects are the nature of the relationship between the data controller and the data subject. What is customary and generally expected in this situation? Also the balance of power between data subject and data controller should be included. Is the data subject obliged under law to provide data? Is the data collection based on a contractual relationship and which freedom does remain to the parties (e.g. to terminate the contract and look for another service provider)? Or is it based on consent and was this consent freely given and to what? Which reasonable expectations concerning limitations of use and confidentiality follow from these circumstances? Generally further use will likely be more limited, the more specific and restrictive the context of the data collection was. Also the transparency of the processing and if it is based on provisions of the law are important elements to consider.
c) The nature of the data and the impact of the further processing on the data subjects. The more sensitive the data, the more strict compatibility will be judged. Both positive and negative impacts have to be taken into account and relevant consequences can vary from well-defined and targeted to more general, unpredictable or widespread. Relevant can be

[150] WP 29, Opinion 03/2013 on purpose limitation, WP 203, 2 April 2013, pp. 11-20.

the way in which the data is further processed, like is it further processed by other controllers, how many people have access, is the data combined with other data, and so on. Also relevant is the availability of alternative methods to achieve the purpose with less negative impacts on the data subjects.
d) The safeguards applied by the controller: safeguards by the controller can prevent negative impacts or limit the risk for such impacts. Therefore they are relevant in the compatibility assessment. Additional technical and organisational measures to ensure functional separation (e.g. pseudonymisation, aggregation of data) or increase the control by the data subject (e.g. improved transparency) can compensate in a limited way a change of purpose or a too vaguely formulated original purpose. To discern what are relevant measures the WP 29 puts forward the basic goals of data security (availability, integrity and confidentiality) and data protection (transparency, isolation, intervenability).[151]

The WP 29 points out that the provision on further processing for historical, statistical or scientific purposes has to be understood in a similar vein. It is not intended as an exception to the requirement of compatibility or as a general authorisation. Also in this case the double test of compatibility and of the presence of a legal ground has to be fulfilled. Rather it permits further processing when the change of purpose is adequately compensated with additional safeguards. This follows clearly from recital 29, which states that such processing "is not generally to be considered incompatible ... provided that Member States furnish suitable safeguards". The recital further states that "these safeguards must in particular rule out the use of the data in support of measures or decisions regarding any particular individual".[152] Therefore the WP 29 also highlights the notion of functional separation. This means that data used for these purposes should not be available to support measures or decisions in regard to individual data subjects. When the data controller can provide adequate safeguards to ensure the further processing will not be used for decision-making about individuals, then such processing can be compatible. Such safeguards will entail all necessary organisational and technical measures to ensure functional separation. Anonymisation techniques can play an important role as safeguard measure. The WP 29 discusses several scenarios, depending on the sort of data needed in the further processing: unidentifiable personal data, indirectly identifiable data or directly identifiable data. In the first case personal data is made unidentifiable by strong forms of anonymisation or aggregation, in which case the data protection framework is not applicable anymore and no extra safeguards are required. But such full anonymisation is difficult to achieve and can destroy the utility of the data. In the second case identification is hampered by pseudonymisation or some form of anonymisation or aggregation, but remains identifiable through combining with other data. Additional safeguards are needed to guarantee the functional separation, like limiting access to and use of the data or 'data silo'-ing. Directly identifiable data may only be used when even partial anonymisation would hamper the purpose of the processing and with other safeguards into place.[153]

[151] WP 29, Opinion 03/2013 on purpose limitation, WP 203, 2 April 2013, pp. 20-27.
[152] 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 October 1995, recital 29.
[153] WP 29, Opinion 03/2013 on purpose limitation, WP 203, 2 April 2013, pp. 28-32.

The WP 29 considers specifically the compatibility assessment and the safeguards needed for further use in big data analytics. It points to its general observations on the factors to consider in this assessment. To determine which safeguards are needed, it considers 2 scenarios. In the first scenario the data controller is interested in detecting trends and correlations on the macro-level but not in the individuals. In the second scenario the data controller is interested in the individuals. For the first scenario the WP 29 points to its considerations on functional separation. When no measures or decisions concerning individuals are made, it is possible that such further processing is compatible. In this case safeguards need to be taken to ensure this functional separation and to guarantee the confidentiality and the security of the data. In the second scenario a specific and informed 'opt-in' consent is in most cases necessary. Such informed consent implies that data controllers will need to disclose their decisional criteria and give access to both the data and the logic behind the processing.

Through this guidance the WP 29 makes clear that the data protection principles can still be applied in a context of big data. It also makes clear how the actual risk involved can be taken into account in the relevant legal assessments. Both the reasonable effort-test concerning identifiability and the compatibility assessment concerning further processing take into account a wide range of factors, including the whole range of safeguards implemented to diminish the risks involved. Such safeguards can be of a technical nature, but also of an organisational or a legal nature. Big data operations with personal data therefore are not made impossible by data protection. Also further processing for new purposes, not foreseen at the moment of collection, is possible but within limits. However, big data operations require a specific attention to the risks involved and to the safeguards which can be implemented to diminish those risks.

4.4.2 Privacy by design

As made clear, data protection law requires to assess the risks to personal data involved in the processing and to take adequate measures to diminish those risks. In a context of big data, with a large amount of interactions involving personal data, it is in practice impossible to make such assessments for each distinct act of processing on each distinct element of personal data. It will be needed to make aggregated assessments of risks and decisions about the safeguards to implement. To make such an overall, aggregated evaluation possible of the risks involved, it is important to develop standard solutions with known capabilities and risks for the whole range of technical, legal and organisational safeguards. Privacy-by-design can therefore be an important approach to develop practical solutions for data protection in big data settings. Privacy-by-design moves a large amount of evaluation and decision-making from the individual acts of data processing to an aggregate assessment in the design phase. Standardisation makes swift risk analysis and auditing of systems possible, thereby lowering transaction costs involved. This both for the big data operator, who has the responsibility as data controller to comply with data protection law, as for the data subject, who as client of big data services has to be able to assess what happens with his data and compare several offerings.

Privacy-by-design originated from another concern. Reducing privacy to using some privacy-enhancing-techniques (PET) can lead to ignoring privacy concerns at an early stage. Often

privacy is just an afterthought, while privacy concerns are ignored during the design of IT systems. In these cases PETs are used to patch the systems when it becomes clear that users have concerns which can inhibit them from using the system. Privacy-by-design has been developed to ensure that privacy concerns are considered at an early stage and privacy objectives are taken up during the design of the system. It will be included as a legal obligation in the new GDPR. In the light of our earlier discussion on safeguards, privacy-by-design can offer an integrated approach to decide about the necessary safeguards. This implies that privacy-by-design may not be limited to technical safeguards, but has to be broadened to organisational and legal safeguards.

The concept of privacy-by-design got developed in a report on PETs in 1995 by the DPAs from the Netherlands and Ontario.[154] Originally the focus was strongly on PETs but it developed under the impulse of the Ontario Information and Privacy Commissioner Ann Cavoukian into a more comprehensive approach. She expressed this approach in 7 principles, stressing the need for a proactive approach to address privacy problems and to early mitigate privacy concerns in the design process and throughout the whole life cycle of the system.[155] These principles reflect very well the intent and the objective of the approach, but have been criticised as being too vague and lacking in how they should be translated into engineering practice.[156] Mainstreaming privacy into design methodologies implies translating these concerns into precise and measurable objectives, which can be included into the requirements for the system under development, and a methodological approach to ensure that these objectives are taken into account and addressed in the design process. Further this mainstreaming is complemented with the development of standard problem scenarios and solutions.

There have been quite some research efforts made to translate privacy concerns and legal requirements into precise and measurable objectives. The WP 29 made use of this work in its Opinion on Anonymisation Techniques when it stated that "'identification' not only means the possibility of retrieving a person's name and/or address, but also includes potential identifiability by singling out, linkability and inference".[157] A state-of-the-art nomenclature of privacy-related objectives and terminology can be found in the work of Pfitzmann and Hansen.[158] This focusses on objectives to be reached with technical measures, but similar work can be done for contractual measures as can be seen in the already mentioned work on cloud SLA standardisation.

---

[154] Office of the Information & Privacy Commissioner of Ontario, Registratiekamer, *Privacy-Enhancing Technologies: The Path to Anonymity*, 2 volumes, 1 August 1995

[155] Cavoukian, Ann, "Privacy by Design: The 7 Foundational Principles", Information & Privacy Commissioner of Ontario, Canada, 2009, https://www.privacybydesign.ca/index.php/about-pbd/7-foundational-principles/

[156] Gürses, Seda, Carmela Troncoso and Claudia Diaz, "Engineering By Design", Computers, Privacy and Data Protection (CPDP), 2011

[157] WP 29, Opinion 05/2014 on Anonymisation Techniques, WP216, 10 April 2014, p. 10-12.

[158] Pfitzmann, Andreas, Marit Hansen, *A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management*, August 2010, http://dud.inf.tu-dresden.de/Anon_Terminology.shtml

A further building block is the research done on privacy patterns. Patterns are reusable solutions for common problems, and can be understood as formalised and abstract representations of best practices. The concept of patterns was originally developed in architecture but became very popular in software design. The development of object-oriented programming incited more profound reflection on how to structure software and led to architectural patterns like the model-view-controller pattern. Now patterns can also be more abstract and reflect design strategies, which do not impose a certain structure but rather an approach to achieve a certain goal.[159] Patterns are also used in security research and recently some, although still limited, efforts have been done to develop design patterns for privacy-related problems and requirements. Some of the approaches discussed earlier, like the k-anonymity model of Sweeney or attribute based credentials can be considered as patterns, as pointed out by Jaap-Henk Hoepman. He developed a set of 8 design strategies based on the legal requirements of the data protection directive: Minimise, Separate, Aggregate, Hide, Inform, Control, Enforce and Demonstrate.[160][161]

Comprehensive design methodologies exist for security and can be broadened to take other concerns into account. We present the LINDDUN methodology as a practical example of such an approach. LINDDUN is a systematic approach to identify privacy risks based on threat modelling and is based on STRIDE, a design methodology for security developed by Microsoft. The advantage of developing a privacy design methodology based on such a methodology for security design is that both can be closely integrated within the software development cycle. STRIDE is an acronym standing for the security threats it includes in its threat model: Spoofing, Tampering, Repudiation, Information disclosure, Denial of service, Elevation of Privilege. Accordingly, LINDDUN defines a range of privacy properties: unlinkability, anonymity & pseudonymity, plausible deniability, undetectability & unobservability, confidentiality, content awareness and policy and consent compliance. From these it derives a range of privacy threats it models and from which it derives its name: Linkability, Identifiability, Non-repudiation, Detectability, Disclosure of information, content Unawareness, policy and consent Non-compliance. These threats have been developed as the negation of security or privacy objectives. The precise formulation of such security and privacy objectives is therefore implied in these approaches. The privacy objectives in LINDDUN are based on the work of Pfitzmann and Hansen, which provides a state-of-the-art nomenclature of privacy-related terminology.[162]

The methodology instructs in 6 steps which issues to investigate and where in the system they can show up. In a first step a graphical representation of the system is defined through a data flow diagram (DFD). In a second step the privacy threats are mapped to the DFD elements. These threats are in the third step further developed into misuse scenarios through threat modelling. As an element of this approach a catalogue of privacy threat tree patterns has been developed and is continuously updated. These trees describe the most

[159] Hoepman, Jaap-Henk, “Privacy Design Strategies”, October 2012, arXiv:1210.6621.
[160] Ibid.
[161] Other research efforts on privacy patterns can be found on http://privacypatterns.org and in Hafiz, Munawar. "A collection of privacy design patterns." Proceedings of the 2006 conference on Pattern languages of programs. ACM, 2006; Hafiz, Munawar. "A pattern language for developing privacy enhancing technologies." Software: Practice and Experience 43.7 (2013): 769-787.
[162] Pfitzmann, Andreas, Marit Hansen, op. cit., 2010

common attacks for each LINDDUN threat category combined with a DFD element type. Each leaf in the tree is examined for its applicability in the system at hand. If applicable it is further documented in a misuse scenario. If not it is documented as an assumption. This makes it possible to easily trace back the needed changes when one of these assumptions does not hold any more. In the fourth step the risks associated with each privacy threat are assessed with an established risk assessment method, like the French EBIOS method. These risks are then prioritised. The fifth step identifies mitigation strategies according to the priorities. The last step translates these mitigation strategies into actual technical solutions (PETs) or into requirements which are taken into account during the further development process.[163]

LINDDUN provides a clear methodology through which engineers can translate general privacy concerns into system objectives and further into actual technical responses answering on practical misuse scenarios. It makes a continuous adaptation possible at several levels (system objectives, threat tree patterns, mitigation strategies) for new technical developments. The limitation is its focus on technical safeguards in the last step, where it could be broadened to legal and organisations safeguards as well. Further this methodology could also be broadened to deal with other societal concerns, like anti-discrimination, once they are translated into precise technical objectives and a catalogue is made of misuse and mitigation strategies. As such this methodology is a useful framework to show how societal concerns in general can be translated from general principles or concerns into actual design practice.

The organisational dimension of privacy by design can be found in privacy impact assessments (PIA). A PIA is an organisational safeguard to early detect risks and address them. The upcoming GDPR will turn this into a legal requirement. The Commission draft foresees a data protection impact assessment when the processing of personal data poses "specific risks to the rights and freedoms of data subjects by virtue of their nature, their scope or their purposes" and defines a range of cases it should take place.[164] The Commission draft provides the minimum content required, like a risk assessment and the measures envisaged to address those risks, and imposes to gather the views of data subjects or their representatives. The European Parliament draft augmented this requirement with a preliminary risk assessment. Depending on the results a full DPIA is required, of which the content is further specified and which has to address the "entire lifecycle management of personal data from collection to processing to deletion". Further the EP draft requires a Data protection compliance review at least each 2 years or when a change in the risks presented by the processing occurs.[165]

---

[163] Wuyts, Kim and Wouter Joosen, *LINDDUN privacy threat modeling: a tutorial, Technical Report* (CW Reports), volume CW685, Department of Computer Science, KU Leuven, July 2015; X., "LINDDUN in a nutshell", https://distrinet.cs.kuleuven.be/software/linddun/linddun.php

[164] European Commission, 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), COM/2012/011 final, 25 January 2012, art. 33(1).

[165] European Parliament, legislative resolution of 12 March 2014 on the proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the

In recent years a lot of research has been done on Privacy Impact Assessments, while also some DPAs have provided guidance on drafting a PIA. As an example we present the guidance offered by the CNIL. The Commission Nationale de l'Informatique et des Libertés (CNIL), the French DPA, has published guidance on how to perform a PIA. In particular, it describes how to use the EBIOS risk assessment method on the processing of personal data. It defines risk as "a hypothetical scenario that describes how risk sources can exploit the vulnerabilities in personal data supporting assets in a context of threats and allow feared events to occur on personal data, thus generating impacts on the privacy of data subjects".[166] The risk level is determined by looking at the severity, or the magnitude of the impact, and the likelihood that the risk occurs.

The CNIL model of a PIA contains 4 main parts. First a description of the processing of the personal data and its context. The next step identifies based on this description the existing and planned controls on the processing. Specific explanation has to be given on how compliance with the legal requirements (like purpose limitation, data quality, the exercise of data subject rights ...) is assured through these controls and how they safeguard against privacy risks. In the third step risks are identified and their level assessed. In this step risk sources are identified, as an answer on the question "Who or what could be the source of risks that might affect the specific context of the processing(s) of personal data under consideration?" Further feared events are identified answering the question "What do we fear that might happen to data subjects?" For each of these feared events the impacts on the data subject are determined and the severity is estimated. Next, for each of the feared events the threats which would cause them are identified, including the risk source from which they originate. An estimation of the likelihood of the threats is made based on the capabilities of the risk source, the vulnerabilities of the data and the current controls. Finally risk and risk levels are mapped. A risk is a feared event with all the threats that can cause it. Using the estimations of severity and likelihood a risk level can be determined. The EBIOS risk assessment method uses 4 levels (negligible, limited, significant, maximum) for severity and likelihood. It makes it possible to visualize the different risks on a risk map with the 2 axes of severity and likelihood and to quickly determine the main risks and how they can be diminished through controls. In the final step the decision is made if the identified controls deal with these risks in an adequate manner or if further remediation is needed. These 4 steps are meant to be part of a continuous improvement process and can be used iteratively till an adequate level of privacy protection is achieved.[167]

Although it shares elements with design methodology, a PIA has clearly a different aim and procedure. Its aim is not to design a system, but the prospective assessment of designs or project plans. It needs to be done before the actual implementation of a project. It is less specific than a design methodology like LINDDUN, as it has no fixed library of threat trees and mitigation strategies. But the PIA can share elements with the design methodology, like the risk assessment, and it entails a broader view on the safeguards, which also include

---

processing of personal data and on the free movement of such data (General Data Protection Regulation), Amendement 127-130.

[166] CNIL, *PIA, la méthode: Comment mener une EIVP, un PIA*, June 2015; CNIL, *PIA, L'outillage: Modèles et bases de connaissances*, June 2015.

[167] Ibid.

organisational measures. A PIA provides a more high-level analysis of privacy risks which can be further analysed in the design methodology. Further a PIA is more aimed at establishing a broad view on potential risks through including the views of stakeholders. As such, it can validate the approach taken through the design methodology or shows the gaps still present, and in an earlier stage show the risks involved with certain projects before going into actual design. It also makes a regular review and update possible, and is therefore not just a stand-alone report but a process accompanying the whole life-cycle of a project. A PIA is not merely a compliance check, but it can also serve the aim of documenting compliance. The PIA further can provide a basis for later review of the implementation and auditing and serve as corporate memory.

Research efforts are made to provide a better link between design and PIA methodologies, like through the Privacy Management Reference Model (PMRM) developed under OASIS (a non-profit industry consortium developing open standards). This model aims to provide standardized concepts and methodologies to operationalise privacy management throughout the life-cycle of a project similar to the practices developed for security. It aims to be "boundary object" or "a construct that supports productive interaction and collaboration among multiple communities", from software and system developers to privacy and risk officers and public policy makers. It entails a conceptual model of privacy management, a methodology and a set of operational services where to map privacy controls.[168] A similar effort which builds on these earlier results is the PRIPARE research project.[169]

To conclude we can observe that privacy by design has developed from a collection of PETs to a comprehensive framework for privacy management similar to security. It ranges from the specification of legal objectives in technical objectives, over standardized threat scenario's and mitigation strategies expressed in design patterns to design and impact assessment methodologies. It shows the range of practices needed to manage societal concerns with technology in a 'by-design' approach. The basic tools are available to deal with privacy in big data practices, but all these areas still need further development and implementation in practice. One further development needed is to include organisational and legal safeguards into the design methodologies, to turn these into an integrated approach to develop socio-technical systems. The privacy-by-design approach also provides a best practice which can be augmented to deal with other concerns as well, like anti-discrimination, or outside the context of personal data, like to deal with confidentiality concerns of companies. Such an integrated design methodology, together with standardisation of the safeguards involved, makes it possible to move a large amount of the assessment and decision-making about privacy from the individual transactions to the design phase of the system. This shift lowers the transactions costs involved and provides a necessary mechanism to deal with privacy concerns in an aggregated manner.

[168] OASIS Privacy Management Reference Model (PMRM) Technical Committee, "Privacy Management Reference Model and Methodology (PMRM) Version 1.0", 3 July 2013; Jouvray, Christophe, Nicolas Notario, Alberto Crespo, Hisain Elshaafi, Henning Kopp, Frank Kargl, Daniel Le Métayer, *D2.1 Best Practice Templates*, PRIPARE, 30 June 2014, pp. 36-42.
[169] http://pripareproject.eu

4.4.3 Transparency and accountability

Till now we looked at how a data controller can apply the data protection principles on big data and came to the conclusion that these principles do not form an insurmountable barrier for big data. Flexible and risk-based interpretations of these principles are possible and a range of technical, legal and organisational elements can be used to fulfil the requirements. However, data protection law does contain also some outdated mechanisms which do not function well in the context of big data. Several elements are based on a transaction-model and do not scale well to a context with a large amount of actors and interactions between them. The data protection framework not just consists of rules on how to treat data. It also sets up a framework for accountability and organises informational control by the data subject through providing it with a set of rights. In this set-up for informational control we find several elements which necessitate individualised action and decision-making, be it from the data subject or the data controller, and which are therefore difficult to aggregate and lead to large transaction costs.

A first example is consent as ground of legitimation. This is not the only ground of legitimation, but one which is very much used. Especially for sensitive information (e.g. health data) the law provides a larger role for explicit consent. And when personal data is used to produce individualised decisions or treatment concerning the data subject (e.g. targeted advertising or commercial offers, risk scores) the WP 29 requests the use of informed consent and an explicit opt-in mechanism. On this issue we can agree with the criticism that consent is not effective any more as a protective mechanism.[170] Often it is not just a hurdle for the data controller but also for the data subject itself. Data controllers can easily shift their responsibility by presenting long privacy notices acting as liability disclaimers and which contain an agreement with a very broad and vague range of further uses. Data subjects are often not in a position to read or even understand such notices and click very fast for agreement to get past it. The experience with the cookie-directive, which introduced a stricter consent and transparency requirement into the e-privacy directive, is a clear example of a dysfunctional consent mechanism. In practice users do not gain much more information and experience the consent mechanism as a nuisance. In other words, in this case the consent mechanism presents larger transaction costs for the data subject and, even when the directive allows for exemptions when for the "sole purpose of carrying out or facilitating the transmission" or when "strictly necessary" to provide a service requested by the user[171], the use of a consent mechanism needs to be more risk-based to be effective as a safeguard. In the context of big data a data controller will often ask a very broadly formulated consent to avoid to have to return too much to the data subject for a renewed consent, which again hollows out the efficiency of consent as a safeguard. We can conclude that in the context of big data the safeguards envisioned in the data protection framework also need a more collective or aggregated solution and that protection has to shift from a transaction model to a liability model.

---

[170] E.g. Tene, Omer, and Jules Polonetsky, "Big data for all: Privacy and user control in the age of analytics.", op. cit.; Solove, Daniel J. "Introduction: Privacy self-management and the consent dilemma." *Harv. L. Rev.* 126 (2012): 1880.

[171] European Parliament and the Council, Directive 2002/58/EC of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications), article 5.3.

Mantelero points out that in the first data protection laws the role of self-determination through consent was not present. These laws came into existence in the context of governments acquiring mainframes, which provided the capability to aggregate information in ways not possible before with paper archives. It provided government with a large concentrated and centralised control of information and data protection laws had to counter this imbalance in power. The main protection was provided by transparency and independent supervisory authorities. Consent did not play a role, as data collection was mostly mandatory and computerised information processing was considered opaque and too difficult to comprehend for citizens to be able to make informed decisions. When with the advent of the PC computers became widely available and made possible the broad uptake of commercial practices based on personal data like direct marketing, consent became more useful as an individual control mechanism. Mantelero points out that in the big data context the old power imbalance and opaque information processing returns and necessitates again more powerful and collective protection mechanisms through supervising authorities. He does not propose to do away totally with consent, but to differentiate between situations where informational self-determination is still feasible and the newer big data contexts where such individual control becomes an illusion because of the power imbalance. For these situations he promotes a larger role for independent regulators, combined with a level of control for individual data subjects through an opt-out scheme. In the context of power imbalances and the impossibility for users to effectively exercise control, the responsibility to make a risk and benefit assessment should shift from individual users to companies under supervision of data protection authorities. Data protection will become more similar to regimes of product security and liability, like the authorisation of drugs and chemicals. This implies also a strict regulation and standardisation of such assessments, comparable to clinical trials. Such regulation needs to happen under control of regulators, as corporate self-regulation does not provide the same independent balancing of interests.[172] In the US context Tene and Polonetsky go in the same direction when stating that "the role of consent should be demarcated according to normative choices made by policy makers"[173], although they opt for a more liberal regime compared to the strongly regulated approach by Mantelero, Mantelero's focus is on the capacity of the data subject to exercise control, but a similar conclusion follows from our analysis based on transaction costs and the need to shift to a more collective and aggregated treatment of interactions and to a liability model. In fact, the draft GDPR already foresees a diminished role for consent by providing that "Consent shall not provide a legal basis for the processing, where there is a significant imbalance between the position of the data subject and the controller".[174] As pointed out by Mantelero, also the strong role for impact assessments in the draft GDPR represent this shift away from individual control.

---

[172] Mantelero, Alessandro, "The future of consumer data protection in the EU Re-thinking the "notice and consent" paradigm in the new era of predictive analytics.", *Computer Law & Security Review* 30.6 (2014): 643-660.

[173] Tene, Omer, and Jules Polonetsky, "Big data for all: Privacy and user control in the age of analytics.", op. cit.

[174] European Commission, 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), COM/2012/011 final, 25 January 2012, article 7.4

The second area where we can find individual transaction-based mechanisms are the data subject rights of access and correction. In this context the picture is more mixed, as the exercise of these rights can also be automated and scaled for a large amount of data subjects. Tene and Polonetsky see enhanced transparency and access rights as a solution to compensate for more limited use of consent. They state that access rights are "woefully underutilized". Their solution for re-balancing the data protection framework in the big data context is an enhanced transparency and access regime to compensate for changing a consent regime to a more opt out-based framework.[175] Enhanced transparency and access is also feasible in the big data context and can compensate for dropping or a more limited use of other safeguards. Problematic are rather those areas where an assessment of the individual's request need to be made, like for the exercise of the right to be forgotten. Google's privacy dashboard provides an interesting example of how transparency and access rights can be put into practice, while also containing automated consent mechanisms (which are in practice opt out mechanisms). Big data can provide granular access with low transaction costs in both ways, so also for the data subject which can fine-tune its privacy settings. As long as the privacy choices can be standardized, such consent and access regime is feasible in the big data context. Still, this example has its limits. First, it is the big data operator who defines the choices and the default. In practice the default is what is most useful for the operator, not what is least intrusive on the user's privacy. Second, none of the big data operators are transparent on how this personal data is used, apart from a vague statement in the privacy notices. It remains opaque how the user is categorised and which of his personal characteristics play a role in the personalised actions, be it personalised search or targeted advertising. Although technical difficulties are not larger than for giving access to the data, such insight in decisional criteria is not given in practice. Further, none of these operators which collect personal data to sell it to other companies give insight to who, nor is it possible to track personal data across the data brokers.

Transparency can be of which data is used, of how it is processed and of how the results are used. Fact is that transparency is rarely very well implemented and when it is it mostly concerns transparency and access to the personal data held by the data controller. The current data protection directive also foresees that the data subject has to be informed of the purposes of the processing and in case of automated decision-making also of the logic involved in this processing. As such the information duties concerning how the personal data is processed and used are relatively limited, although the WP 29 makes clear the data subject has to be informed of the decisional criteria involved in the processing.[176] Automated decision-making is also weakly regulated in the current directive. The GDPR will probably strengthen these informational duties, especially in the case of profiling and automated decision-making, but how remains a point of discussion in the legislative negotiations. We will return to this point when we discuss anti-discrimination, where we will see that transparency about the use of data is necessary to effectively prevent discriminative treatment as a result of big data processing. Transparency on how data is processed and used is a necessary part of improving accountability in the context of big

[175] Tene, Omer, and Jules Polonetsky, "Big data for all: Privacy and user control in the age of analytics.", op. cit.
[176] WP 29, Opinion 03/2013 on purpose limitation, WP 203, 2 April 2013, pp. 45-47.

data. It needs the development of standardised auditing procedures and assessment tools. Both data protection authorities and equality bodies have a task to cooperate on developing the accountability framework for big data processing.

The practice of Google also makes clear where the exercise of data subject rights makes an aggregated and automated treatment difficult to impossible. The exercise of the right to be forgotten requires an individualised balancing of interests, the privacy interest of the data subject versus the public interest in the information. The Google practice can be seen as one of the best practices at hand and shows it is not impossible to treat a large amount of requests, but it also shows the large cost involved as the individual treatment obliges to build a large administration to deal with requests to withdraw search results.
On the other hand, the EUCJ decision on the right to be forgotten also is an example how this balancing of interests can be modulated in the context of big data. A positive answer to requests to change the original source in newspapers or official publications would come down to falsification of the historical archive. However, limiting the access at search engines or news aggregators blocks the re-use of this data and prevents this information to become a network good and its resulting propagation. In this sense is concentrating the effort to deal with the right to be forgotten at search engines also a rational choice, although it remains a costly practice.

Conclusions

The higher visibility and penetration of boundaries by data flows also results in privacy concerns for individuals. Data protection law presents an important protective mechanism, but has some problems to deal with big data: Big data conflicts with the purpose limitation and data minimisation principles. The network effects of big data make anonymisation techniques unreliable. Further mechanisms like consent and data subject rights reflect a transaction model, which leads to high transaction costs for big data practices.

The WP29 defends the applicability of data protection on big data and points to the risk-based elements in this framework. The purpose limitation principle does not prevent further processing for other purposes, as long as these purposes are not incompatible. The compatibility assessment includes assessing technical and organisational safeguards implemented by the data controller. The most contested technical safeguard is anonymisation, as all methods fail to guarantee anonymity. Despite this, it remains an important safeguard, in particular in combination with other technical and organisational safeguards. The whole of these safeguards can be assessed with the reasonable effort-test and the compatibility assessment.

Privacy-by-design provides an important approach for mainstreaming privacy in the design process. The original principles remained vague but steps are now taken to translate into actual technical development processes. This approach concerns not only a technical approach, but also organisational tools like risk assessments and data protection impact assessments. Although further development remains necessary, privacy-by-design has developed into a practical approach, which can also serve as an example to address other societal concerns.

More problematic is the application of the informational control model through data subject rights to big data. Mechanisms like consent or the need to make an individual assessment of the application of the right to be forgotten lead to high transaction costs. Also here is a shift to a liability model needed by giving a more important role to the regulators. Transparency is technologically feasible and scalable, although rarely implemented fully. Standardised auditing and evaluation tools controlled by the regulator can be a solution.

5 POLITICAL EXTERNALITIES

The political externalities uncovered in the case studies related mostly to political economics. Mainly two issues came forward: on the one hand the relation between public sector or non-profit organisations and the private sector, on the other hand the fear to lose control to actors abroad, and in particular US-based actors, which sometimes translates in protectionist attitudes and requirements to store data within national territories. Depending on the case study, these issues could overlap and were often framed differently. Another political externality, which was only signalled in a few case studies but which remains part of the public debate, is the threat of surveillance and political abuse.

5.1 THE RELATION BETWEEN THE PUBLIC AND NON-PROFIT SECTOR AND THE PRIVATE SECTOR

Potential problems concerning the relation of the private sector with the public and non-profit sector can be summarised as the fear for lock-in and dependency on dominant private players and the fear for rent-seeking by the private sector on public investment.

The fear for lock-in and dependency was raised in the crisis, culture, environment and smart city case study. The negative effects were perceived differently and ranged from economic to political effects of losing control and political abuse. In the crisis case study choosing for partnerships and cooperation with big private players is seen to be advantageous in terms of cost efficiency and implicitly also presents a diversification of funding. But the interests of these partners have more to do with brand promotion and are therefore more divergent compared to those of public funders. As a result the private partners are seen as less reliable, in particular on longer term, and opting for technical solutions which bind to such partner can in the end turn out negative. In the environmental sector the concern was raised that over-dependency on centralised services by large private players also gives them the power to implicitly set standards. In the smart city case study the concern was a mixed economic and political one and had to do with the current dominant position of US-based technological companies and how this stifled prospects for a European data economy. In the cultural sector case study the fear for dependency was also a translation of the political concern to lose control.

5.1.1 Lock-in and dependency in value chains

The problem of lock-in does not concern only the public and non-profit sector but is a more general problem linked with keeping markets competitive. An interesting approach to evaluate the concerns for lock-in and dependency can be found in the research done on power in global value chains. This research originates from development economics and concerned mostly commodity value chains, but some of its insights can be used fruitfully in this context as well. Here we give a summary of this approach. In the next section we apply it to big data value chains.

Gereffi, Humphrey and Sturgeon identified three variables which play an important role in determining how global value chains are governed:

1. The complexity of information and knowledge transfer required to sustain a particular transaction. When only price and basic product information is needed to perform the transaction, this complexity is low. When compliance is needed with extensive product and process specifications, such complexity is high.
2. The ability to codify this information and knowledge. The better such information can be codified, the easier it can be transmitted and the need for transaction-specific investment gets reduced.
3. The capabilities of suppliers in relation to the requirements.

Based on these variables they identified five forms of value chains found in practice:
1. Markets: Asset- or transaction-specific investment can be avoided when the information complexity is low and suppliers can fulfil the demand without much input from buyers. In this situation buyers and suppliers can easily switch and market governance will apply.
2. Modular value chain: Information complexity is high, but this information can be adequately codified. Codification can result from standardisation. Products or services are offered with a modular architecture. This makes it possible that transactions take place with a limited exchange of information and without the need for continuous monitoring, feedback or coordination. This simplifies transactions and no large and specific investment in a supplier relation is needed, which keeps the cost of switching low.
3. Relational value chain: When on the contrary this information cannot be codified, more frequent information exchange and coordination is needed. This makes transaction costs high and requires a high investment in the buyer-supplier relation. These specific costs are specific to this relation and are lost when switching suppliers. This situation also presumes the availability of capable suppliers, which implies that outsourcing remains more cost-efficient compared to in-house development of these capabilities.
4. Captive value chain: when both the complexity of the information and the ability to codify are high, but the capabilities of suppliers (or clients) are low this will require a lot of intervention, in terms of control and support, of the lead firm. This situation enables the lead firm to lock-in suppliers and keep them dependent. Generally the weak partner is confronted with high switching costs and therefore held 'captive', while the lead firm has a power position through which it can capture a major share of the value.
5. Hierarchy: when information is complex, but it cannot be easily codified and the capacity of external partners is low, lead firms are forced to develop the value chain concerned internally. The need for transaction-specific investment is internalised.

Gereffi et al. point to the dynamics in chain governance. Suppliers can acquire the necessary competences and change the governance of the value chain from captive to a modular type. The better power balance in this situation enables them to easier switch between value chains and to capture more of the value. Technological changes also can influence chain governance. When new technologies make the transaction more complex, the chain can evolve from a market situation to more modular, relational or captive relations depending on the codification and capabilities. On the other hand, when transactions become less complex through standardisation or improved access to certain technologies the relations can evolve to modular or market situations.[177]

[177] Gereffi, Gary, John Humphrey, and Timothy Sturgeon, "The governance of global value chains", *Review of international political economy* 12.1 (2005), pp. 78-104.

5.1.2 Lock-in and dependency in the big data value chain

This general account of power in value chains we now apply the big data value chain. The earlier observation that the relations between partners shift from an exchange of goods to a more continuous provision of services shows that transactions involve more complex information transfers. Therefore, earlier modular supply chains or market situations run the risk of becoming captive relations when technological innovations by one firm enable it to capture a dominant position and other actors do not have adequate capabilities to catch up and preserve their freedom to switch partners. The current dominance of US big data players can be explained from this analysis. Early stage technology developments, when highly competent suppliers are not available or only very limited, lead to in-house development. Such in-house development is expensive but can turn into a competitive advantage when it makes it possible to offer new services and therefore to enter new value chains. This is clearly the case with the big data capacities developed by current dominant players. Amazon developed from Internet book store to a major cloud service provider. Google turned its search engine into a marketing tool, while also extending its range of services (e.g. Google maps, YouTube). This enabled them to disrupt old value chains and build new ones from a dominant position.

An important element is that the presence of positive network effects strengthens the development of monopolies or dominant players and raises the entry barrier for later entries. Disruptive innovation in a network environment therefore often leads to monopolies or dominant positions. This consequence of positive network effects has been visible in 'standard wars' since the 1980s. Once a standard could get a dominant position the market tipped in its favour. Other players were driven out, even when technologically superior, as no buyer wanted to run the risk to end up with technology which was incompatible with most of the other users. Standard wars are all or nothing games. The ability to codify, or standardise, is therefore a key instrument. When a company is able to set a standard, it locks the market to its technology. When this standard is linked to its intellectual property, it also acquires the position of gatekeeper which can decide on the capability of other actors in the value chain. This explains the attractiveness of pushing a private standard. On the other hand, for who is not capable to do so on its own it is attractive to make sure he is part of the winning team by setting the standard in cooperation with others or to keep the market open by adopting an open standard. Which attitude to take towards standards is therefore an important element of a competitive strategy. Depending on their power position and their place in the value chain (at the spot where most value gets captured or not), companies will push a private standard or cooperate to adopt an open one. Apple keeps its iPhone-environment closed, while Google provided Android as an open standard as an entry strategy in the mobile market and because its revenue is not linked to selling phones but rather to the marketing revenues delivered through the wide accessibility of its services.

5.1.3 Measures to avoid lock-in

Here we will present a range of measures through which lock-in can be prevented: collective standards setting, open source and open protocols, data portability and open data.

Avoiding lock-in requires keeping a competitive environment with the presence of many players at the different stages of the value chain. In the analysis of Gereffi et al. the key variable between modular and captive value chains are the capabilities of the suppliers which enable them to switch value chains and avoid capture by a lead firm. In this context it implies preventing the use of technical standards or platforms as gatekeeping tool. This can be done by collectivising standard setting and by turning closed platforms into open protocols and open source software.

Standardisation makes it possible to turn captive value chains into modular value chains by codifying information transfers. Open standards lower the entry barrier into the stage of the value chain in which the lead firm is active. It raises the capabilities of the actors in the other stages of the value chain as they can switch from value chain without losing asset-specific investment. This enables new players to easier compete with the lead firm. E.g. in the long run we can expect the cloud service market to open up through standardisation. No positive network effects prevent other actors to enter. Only economies of scale are at play and diminishing prices of equipment caused by technological improvements will lower the barrier to entry.

Lock-in will present more of a problem in the data usage stage, where data-driven services can more easily acquire and sustain a dominant position. Lock-in is in this case caused through concentration of data holdings and having gatekeeping power for access to that data. Such data holders can also fully profit from positive network effects caused by open data, while blocking or limiting such effects for others. Open data can be a strategy to avoid data monopolies, but it raises questions about rent-seeking by others. The main value from open data will be captured by those who have already a dominant position through their data holdings and it can strengthen their position. New players will only be able to capture value from leftover niches as long as open data holdings do not substantially lower the barriers to entry. E.g. providing time tables of public transport as open data enables a range of small players to develop apps and web-services, but it also strengthens dominant players like Google maps which can further broaden the use of their already established service in new markets and drive the small players out. A defensive strategy, in which players keep their data closed and develop their own app or service, will not be able to open up the market. Only a collective approach, which involves enough actors to overcome the barrier to entry raised by the positive network effects in favour of the dominant actor, can change the market environment substantially. E.g. the Android open source-strategy by Google, the collaboration to set the GSM or the DVD-standard. For data holdings this would mean developing common data standards, which assures sufficient data quality and making it available as a whole. The Europeana-project could be seen as an example of this, while also good efforts have been made concerning geodata. Other efforts for public data could be developed as part of the European Interoperability Strategy or open data efforts. But also private, sector-wide efforts are possible, although rare or non-existent in practice.

Another measure to avoid lock-in is to assure data portability. Data portability has been raised in the discussion on data protection, mostly in the context of social media, and is included in the draft GDPR. The right to data portability is in this context a specification of the right to access to personal data, and allows the data subject to receive its personal data

"in an electronic format which is commonly used" for further use[178], which can include the transfer to a competitor for similar services. On the other hand, data portability serves more the objectives of competition law than of data protection.[179] A generalised right to data portability would be an important measure to keep the market in data services competitive. It prevents lock-in by putting the obligation on the service provider to include the possibility to obtain the data in a commonly used format. Such providers are obliged to ensure a minimal form of interoperability between them and thereby also lowers the barrier to entry to the market for their services.

5.1.4 Open data and rent-seeking

As mentioned before, the effect of open data can be ambiguous. This relates to the observed fear for rent-seeking on public efforts by private actors and concerns in particular investments in open data. Open data is promoted as a way to unlock the value present in this data. It is based on a market expansion effect resulting from the positive network effects associated with data when it is made interoperable. This reasoning is similar to the analysis of open source software provided by Yochai Benkler in The Wealth of Networks.[180] But it tends to forget that the individual incentive to open the source code or data does not have to be the same as the overall societal value of opening. For source code this incentive and the societal value are more similar, as it makes it possible to capture benefits from the collective investment in correcting and further developing the code. The individual contribution is less compared to writing the software on its own. However, opening data does not make your data better while also the investment to make this data interoperable can be substantial. Keeping your data closed when other sources are open can even deliver you a competitive advantage. In other words, costs and benefits are not so evenly distributed. This leads to a strong focus on governments opening data and very little on such efforts by the private sector. Governments can capture some direct effects if they see this as a form of outsourcing. Governments are well-placed to collect information, but often not so good in turning these into accessible information resources and services. Making data open can lead to private efforts to develop such services, which the government can use for free or at a more competitive price compared to licensing the data to one player or develop the service in house. E.g. providing business registers as open data makes possible the development of several private services based on this data source and the commercial courts or other governmental services can obtain these services at a lower and competitive price, compared to the price of the service when licensed exclusively to an individual actor which later sells its services back to a range of governmental services. But this is valid only in a limited amount of situations. In general governments open their data as a resource as part of their economic policy (or as part of an effort to make government more transparent) and hope to capture indirect effects through the development of a data economy. The private

[178] European Commission, 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), COM/2012/011 final, 25 January 2012, article 18.

[179] Tene, Omer, and Jules Polonetsky. "Big data for all: Privacy and user control in the age of analytics.", op. cit.

[180] Benkler, Yochai. *The wealth of networks: How social production transforms markets and freedom.* Yale University Press, 2006.

sector has much less incentive to open its data, while it is eager to capture benefits from public open data. To conclude, the bazaar model from open source software where everyone copies from each other and contributes in order to use a collectively improved product, does not hold for open data. And this results in the dependency on public funding, signalled in several case studies.

To evaluate if open data achieves its policy objectives or leads to rent-seeking, these other policy objectives have to be defined and measured. This can be compared with the cost through subsidies in a cost-benefit analysis. In practice only subsidies are available as a revenue model for public open data. Also other indirect revenue models can play a role, but are more suited for private sector purposes. The LOD2-project listed several revenue models available for linked data. Direct revenue models are subsidies, licensing and subscription, indirect revenue models are advertising, commission, attracting traffic and branding.[181] Licensing and subscription point to a limited access model in which access is given upon payment. That leaves only subsidies available for open data as a direct revenue model, while also the indirect revenue models can be used for open data. Evaluating rent-seeking leads therefore to a macro-level evaluation if the aim is to stimulate a data economy. When there are a lot of potential users, open data is the preferred option and its return will be indirect through the realisation of the policy aims. However, when the data source is used by only a few companies to make commercial applications, licensing the data can be a method to share the revenue and finance the investment in open data. When the government is an important user of these applications, the cost of such use has to be included in the evaluation.

To conclude we can say that the presence of positive network effects in the market for data and data services can easily lead to lock-in and captive value chains. Key elements determining if captive value chains can be turned into more modular ones are the ability to codify the information exchange and the capability of other actors to control or participate in this codification and retain their capacity to switch between value chains. Tools to prevent lock-in are therefore collective standardisation processes, open source and open standards, open data and data portability. However, open data can have ambiguous effects, as it can also strengthen existing dominant positions. Although in general open data policies have a market expansion effect, a macro-level evaluation has to be made to check if it results in or strengthens dominant positions. In this case capturing the monopoly rent through providing the data through licensing can be a better solution.

5.2 LOSING CONTROL TO ACTORS ABROAD

The fear for dependence of other actors also translates in a problem for governments. Again this can be an economic concern and taken up in economic and trade policy. In this case it can lead to protectionist measures. Such measures are negative seen from a macro-level perspective, but can be rational from an actor perspective aiming to prevent the 'creative destruction' of its economic base or rent-seeking and to retain its economic independence and control over its resources. It can also be a proper political concern to keep its ability to

[181] Pellegrini, Tassilo, Christian Dirschl, Katja Eck, *Deliverable 7.5: Asset Creation and Commercialization of Interlinked Data*, LOD2, 29 July 2014

regulate and control intact. The digitalisation and the advent of the Internet has affected strongly the capacity of states to control and regulate societal activity. Big data, when linked to the dominance of US companies, can further erode this capacity for European governments and transfer implicitly the authority to regulate such activities to the US government. On the other hand, big data can also be a tool to restore such capacity to control, which by consequence also raises questions concerning potential abuse and how to safeguards against this.

In D2.1 we raised how the technical convergence and reordering of services, which result from the use of cloud computing and big data processing in business processes, can also lead to jurisdictional problems. Cloud computing makes the shift possible of in-house activity to services provided over the Internet, leading to complicated architectures with several partners which can be located worldwide. In this context borders do not function well. We noticed that legislators, out of fear that Internet-related activities escape regulation, are tempted to react by including light mechanisms in their legal frameworks to trigger them into application. This results in practice in an extra-territorial reach of legislation and can result in an inflation of applicable laws and in conflicts concerning the applicability of laws.

In the case studies this problem was not raised as such. Case study partners were rather confronted with laws requiring to store data within national borders, like in the energy and the culture case study. Instead of giving legal frameworks extra-territorial reach, these efforts try to resurrect borders and national control. They can pose a legal barrier for interoperability. But often these laws do not prevent copying and making the data available abroad, and can therefore be seen as a relative minor measure to prevent a loss of control over the data. Main negative effect is in this case the extra cost associated with fulfilling these national requirements. Further the problem was raised in the crisis case study as unintentionally resulting in the data being subject to US law, which illustrates the problem of loss of control. In this case it could easily lead to conflicts between laws, although they appeared not in practice (yet). But the recent decision of the EUCJ in the Schrems-case, which cancelled the Safe Harbour-decision of the European Commission, shows that for organisations active in the EU and using US infrastructure such problems can become more acute.

A best practice to deal with legal conflicts and the extra-territorial reach of legislation, while also retaining the capacity to regulate, would be legal harmonisation. Instead of individual efforts to retain this capacity by expanding the reach of its legislation, harmonisation of legal frameworks collectively restores the capacity to regulate at the international level. Such harmonisation can take a lot of shapes, ranging from a full-blown treaty to mutual recognition mechanisms. The data protection framework includes such mutual recognition mechanism and the Safe Harbour decision of the European Commission was an application of it. However, legal harmonisation is often easier said than done, and requires also the alignment of a range of interests and values. States will balance these interests with the benefits from legal harmonisation, including by comparing the loss of control associated with both situations. States which retain a strong capacity to act unilaterally will often be reluctant to engage in harmonisation or will only do so when it is compatible with this capacity.

The Schrems-case is a clear illustration of this problem. As said, the Safe Harbour decision of the European Commission concluded that this US framework provided an adequate protection of personal data comparable with the protection offered in the EU. The EUCJ decided otherwise, due to the lack of legal control on the access of security services to personal data held by companies, as became clear through the revelations about the NSA by whistle-blower John Snowden. The EUCJ pointed to the fundamental rights developed through the data protection framework.[182] Implicitly it made clear these are interests which need to be respected in any case and cannot be traded in negotiations. On the other hand, the US sees the NSA activities as central to its security and remains reluctant to limit itself. A solution to this conflict lies in the negotiation of a new Safe Harbour-agreement or similar, as made clear by the WP 29.[183] The EUCJ has made respect for the EU fundamental rights a bottom line for such agreement, which includes a refusal of massive and indiscriminate surveillance. Respect for this bottom line implies a limitation of the current NSA practices as revealed by Snowden. In other words, this will be a hard nut to crack. Similar legal conflicts can arise in other areas. E.g. the CNIL tries to enforce the application of the right to be forgotten by Google on its search results worldwide. This is in conflict with the larger constitutional protection of free speech in the US. Similar problems can occur when more censor-minded states try to enforce their laws worldwide, which can lead to similar conflicts with free speech protection in the EU. General solutions or 'best practices' are difficult to propose.

To conclude, both the use of unilateral protectionist measures and the extraterritorial application of national legal frameworks to restore or retain the regulating capacity of the state can lead to legal conflicts or to barriers for big data practices. Although often difficult in practice due to conflicting interests, international legal harmonisation is the best approach to address these problems.

5.3 POLITICAL ABUSE AND SURVEILLANCE

Lastly we consider the issue of surveillance and the fear for political abuse of big data. This was only raised in 2 case studies (crisis and environment). In the smart city case study fear for manipulation was raised as a concern towards commercial actors, but in the validation workshop it was pointed out that this could concern political abuse as well.

Big data can be used to restore governmental control with new means. In fact, the loss of control through the Internet or globalisation is counterbalanced by the efforts of governments and law enforcement agencies to develop and implement new means of control. The new availability of data is used to create new visibilities of behaviour of citizens or companies. Such big data practices can be very intrusive and lead to a heightened surveillance.

Across the EU governments are setting up projects to make a range of data available for data mining with the purpose of combating social or fiscal fraud, criminal activity and so on.

[182] EUCJ, Maximilian Schrems v Data Protection Commissioner, C-362- 14, 6 October 2015.
[183] WP 29, Statement of the Article 29 Working Party on the Schrems judgment, 16 October 2015.

E.g. in the Netherlands mobility data, like the data from parking meters or about public transport use collected with smart cards, is regularly demanded by the tax authorities. Such fine-grained data reflects daily behaviour patterns and amount to a very intrusive and indiscriminate surveillance. The EUCJ made clear in its decision on the data retention legislation that indiscriminate surveillance, without any specific suspicion, is not compatible with the fundamental right of privacy. On the other hand, big data can play an important and constructive role in law enforcement.

The main question is how to develop new safeguards and balances. It is impossible to present a full-blown proposal for such new safeguards in this context, but some directions can be given. In general it can be observed that the trend has been to limit or cancel legal safeguards when big data processing was incompatible with the existent legal framework. We can also observe that the current data protection framework provides no real safeguards, as it does not apply on law enforcement and its safeguards are in general put aside when the information is requested for law enforcement purposes. Therefore a clear need exists to develop new legal safeguards, which are effective within the new data-driven law enforcement practices. Profiling needs it proper regulation, also in the law enforcement context. In particular, new methods to formulate reasonable suspicion have to be developed which can trigger the access to the data. Purely model-based suspicions have to be avoided, as they risk to be rationalised prejudices. Instead data-driven investigations, and suspicions which trigger them, still need to be linked to concrete and factual elements. A relevant decision in this context is the German Rasterfahndung case, which can be translated as data screening, decided by the German Constitutional Court on 4 April 2006. This decision was induced by the complaint of a Moroccan student against a large-scale data mining operation searching for possible terrorist sleeper cells after the terrorist attacks of 11 September 2001. First a dataset was made of persons that fulfilled a set of criteria, which included being of Muslim faith. In the state Nordrhein-Westfalen this dataset concerned about 11000 people, distilled from a set of 5200000 people. Secondly this dataset was compared to find suspicious matches with a range of other databases, which contained data concerning between 200000 and 300000 people. This screening operation was followed by other investigation measures, but in the end gave no results. The description let us suppose the operation did involve querying of databases, but probably made no use of data mining algorithms. The German Constitutional Court declared the screening operation unconstitutional. The Court pointed out that the distinct steps of the screening operation were all interferences with the informational self-determination of people against who no suspicion was present. Such interference could only be made proportional in limited circumstances. More specifically a concrete danger or threat has to be present and this determination has to have some factual ground. General threat assessments are not sufficient, as this would lead to an unrestricted competence and searches “ins Blaue hinein” (fishing expeditions).[184] This case concurs with the decisions on data retention made by the EUCJ and several Constitutional courts, but decides about a specific operation and not a legal framework for data collection. It makes clear that data-driven law enforcement may not become model-driven law enforcement without the presence of factual elements with which enable the reasonableness of the model can be assessed. Such reasonableness remains the basic element to assess the proportionality of the access and use of personal

[184] Bundesverfassungsgericht (BVerfG), 1 BvR 518/02, 4 June 2006.

data for the purposes of law enforcement and combating fraud. Legal limits to the interoperability of data are therefore also crucial as safeguards for fundamental rights in the context of law enforcement.

We can conclude that the use of big data for governmental control and law enforcement also necessitates the development of new legal safeguards, which limit and specify the circumstances in which governments may collect and use the wide range of data sources to which they have access.

Conclusions

To conclude we can say that the presence of positive network effects in the market for data and data services can easily lead to lock-in and captive value chains. Key elements determining if captive value chains can be turned into more modular ones are the ability to codify the information exchange and the capability of other actors to control or participate in this codification and keep their possibility to switch between value chains open. Tools to prevent lock-in are therefore collective standardisation processes and open standards, open data and data portability. However, open data can have ambiguous effects, as it can also strengthen existing dominant positions. Although in general open data policies have a market expansion effect, a macro-level evaluation has to be made to check if it results in or strengthens dominant positions. In this case capturing the monopoly rent through providing the data through licensing could be a better solution.

The fear for dependence of other actors also translates in a problem for governments. This can be an economic concern and lead to protectionist measures to retain its economic independence and control over its resources. It can also be a proper political concern to keep its ability to regulate and control intact. States can take measures to keep data at home and to give its jurisdiction extra-territorial reach, which can lead to legal conflicts. In this case legal harmonisation is the logical best practice, which can range from a full-blown treaty to mutual recognition mechanisms. But it is often not easy to realise as other interests can be at play, like national security versus fundamental rights in the Schrems-case.

Big data can also be used to restore control and the ability to regulate, but such big data practices can be very intrusive and raise the fear for heightened surveillance and political abuse. Also in this case an urgent need to develop new safeguards is present, like a proper regulation of profiling and data screening by law enforcement and other public agencies.

6 RECOMMENDATIONS

The case studies performed by the BYTE project uncovered a range of economic, social and ethical, legal and political externalities. This deliverable presents an evaluation of these externalities and of best practices to address them. Based on this evaluation a set of recommendations can be formulated for each category of externalities.

Economic externalities

Eight best practices to deal with economic externalities and capture the benefits of big data were identified:

Best practices	SMEs	Large corporation	Research institutions	Public agencies
Public investments in infrastructures				X
Funding programs for big data				X
Public investments in open government data				X
Persuade "big actors" to release some of their data		X		X
Promote big data in education policies				X
Look for interesting data sources	X	X	X	
Seek opportunities in new business models	X	X		
Create partnerships around data	X	X	X	X

Governments and public institutions can diminish the negative impact of scarce resources by public investment in infrastructures and funding research and innovation programmes for big data. They can also fuel the data economy by making more data available through investments in open government data and by persuading "big actors" to release some of their data. This last practice concerns large corporate actors, but public institutions can create incentives for corporate data sharing. Another important task of government is to include attention for big data and data skills in education. Education policies have to address the current scarcity of data scientists and data analysts, but also promote the inclusion of data skills in a range of educational programs.

The last 3 practices concerns changing the mind-set of corporate actors, both large and SMEs, and research institutes to better value opportunities concerning data and to adopt big data practices. Opportunities concern both the creation of new data-driven business models and the exploitation of existing data sources. Public authorities, corporate actors and scientific research institutes can all participate and create new partnerships around data and make new big data use cases possible.

Social and ethical externalities

Positive social benefits can be captured with the same best practices as recommended to address economic externalities. The evaluation based on the interaction perspective showed interoperability as the key factor in enabling big data practices. Investment in the

several dimensions of interoperability (semantic, technical, legal, organisational) helps to make big data practices possible and capture the associated benefits.

Anti-discrimination policies have to address the risk for unintended (or masked) discrimination resulting from data mining. Technical methods to discover and prevent discrimination in data mining need to be further developed and taken up in an 'anti-discrimination-by-design' approach, similar to and embedded in 'privacy-by-design'. The equality bodies will have to take up the task to drive the development and promotion of anti- discrimination-by-design as part of their mainstreaming efforts. Further, they will have to develop a transparency and accountability framework which will be partly based on anti-discrimination legislation but also on the data protection framework. For this they will have to cooperate with data protection authorities. Cooperation between DPAs and equality bodies to share expertise and build a common agenda seems necessary to give anti-discrimination in data mining and big data adequate attention. Lastly, the equality bodies will have to engage in a prospective societal and political debate on the 'new discriminations' made possible by big data. This will have to lead to the development of new norms, be they social, professional or legal.

Legal externalities

Legal frameworks have to be reviewed in order to restore the optimal balance between the different objectives and interests concerned in the context of big data. Regulations need to be adapted to make sure they remain effective in a context with a growing amount of interactions and data flows between actors. Legal mechanisms based on individual transaction- or individual control-models need to be substituted with aggregate or collective mechanisms. Further, the development has to be stimulated of 'by-design'-approaches, in which legal objectives are translated in technical requirements and taken into account during the design of technical systems.

Copyright and database protection: The benefits of big data are best captured by diminishing the barriers posed by IPR on interoperability. This can be done by limiting the extent of exclusive rights on data, by providing larger exceptions in the copyright framework for large scale uses of data like data mining and by abolishing the sui generis protection of databases.

Trade secrets: In order to address confidentiality concerns of companies when participating in big data practices and to restore their gatekeeping capacities, the adoption of a legal protection of trade secrets is recommended. Within such a framework stakeholders and standard bodies have to develop standardized solutions and a toolbox of legal, organisational and technical means which can be used to fine-tune data flows by organisations.

Privacy and data protection: The principles of the data protection framework can be implemented in a risk-based approach without limiting the extent of their protection. But in such a risk-based application it is important to evaluate the actual risks associated with the overall set-up of technical, legal and organisational safeguards, instead of assessing the theoretical risks associated with anonymisation techniques or other technical safeguards. To

make such overall evaluation possible, the development of standard solutions with known capabilities and risks for the whole range of technical, legal and organisational safeguards needs to be promoted. Further, the privacy-by-design approach needs to be further developed and broadened from a purely technical approach to include also legal and organisation safeguards. Lastly, individual transaction-based elements of the data protection framework, like consent, need to be substituted by more aggregated and collective mechanisms, which includes a strengthened role for DPAs.

Political externalities

Lock-in and dependency on dominant private players can be avoided by preventing the use of technical standards or platforms as gatekeeping and control mechanism. Tools to prevent such private gatekeeping are collective standard setting (instead of private), the use of open source, open protocols and open data, and data portability. The effect of open data can be ambiguous and strengthen the position of already dominant players. Therefore the result of releasing data has to be assessed in a specific cost-benefit analysis.

Unilateral protectionist measures or extraterritorial application of national legal frameworks in order to restore the regulating capacity of the state can lead to legal conflicts or to barriers for big data practices. Although often difficult, international legal harmonisation is the best approach to address these problems.

Big data can itself be used to restore governmental control capabilities, but with risks for human rights infringements and political abuse. New legal safeguards need to be developed to restore the balance between citizens and state authorities.

Conclusion

This comprehensive evaluation of societal externalities of big data shows a broad agenda for policy-makers consisting of updating legal frameworks, the promotion of big data practices through public investments, and enabling policies and an active policy to keep markets open and competitive. Regulators and stakeholders also have an important role in developing tools and approaches to mainstream societal concerns into the design process and the big data practices.

7 GENERAL CONCLUSIONS

In this deliverable we have presented a comprehensive evaluation of economic, social, ethical, legal and political externalities of big data. It reveals a broad agenda for policy-makers to address these externalities, consisting of updating legal frameworks, the promotion of big data practices through public investments and enabling policies and an active policy to keep markets open and competitive. Regulators and stakeholders also have an important role in developing tools and approaches to mainstream societal concerns into the design process of and the implementation of big data practices.

In the first part we have analysed the economic effects and externalities associated with the use of big data in the BYTE case studies, as well as in other relevant big data initiatives. The economic externalities treated are categorized into operational efficiency, innovation, new knowledge, business models, and employment. In order to diminish the negative economic impacts and to augment the positive effects, we have devised a set of best practices. These best practices are not only useful to capture positive economic externalities, but appeared useful for positive social externalities as well. Governments have an important role, as most of these best practices are directed towards policy makers and public agencies. Generally governments and public institutions can promote big data practices through public investments and enabling policies. These best practices point to the need for governments to develop a comprehensive big data policy. The best practices concerning corporations point mainly at a need to change the mind-set in order to perceive the opportunities of big data. However, this also necessitates attention for adequate legal frameworks to create legal certainty and diminish negative effects, which now cause distrust and reluctance towards big data.

To get a more in-depth view on the negative the economic approach of externalities, which makes an evaluation on macro-level, is too limited. An analytical frame, which takes into account how big data affects the interactions between different actors, individuals or organisations, proved useful. Through its impact on these interactions, big data can also unsettle current practices and legal frameworks dealing with these interactions and thereby create the negative externalities observed in the case studies. The following observations characterise the impact of big data on these interactions. The capacity to collect, link, process and analyse data on a large scale has turned data into a network good. Positive network effects are an incentive for business models which demand more interactions, in the form of data flows, with other actors. These larger data flows penetrate organisational boundaries and affect their gatekeeping functions. They result in a higher visibility of actors and enable faster and targeted reactions towards them. Data-driven business processes also imply a shift from limited, time-bound transactions to regular or continuous services. The externalities uncovered in the earlier empirical work can in general be explained from these changes in interactions.

Based on this analysis, we can draw some general conclusions on what are best practices to deal with negative externalities of big data.

First we have to check if the balance struck between different objectives and interests in the current practice or legal framework is still delivering an optimal result. If not, it will need

adapting. Such adaptation can lead to different conclusions depending on the objectives of this practice or legal framework. E.g. for copyright a limiting of the protection proved more effective to reach the overall goal of stimulation innovation. Such a solution was not useful for data protection, where we deal with 2 incommensurable objectives of economic benefit and privacy. In this case adapting the framework to make it more effective strikes a better balance.

Second, legal mechanisms based on individual transactions or individual control lead to high transaction costs in a context of a larger amount of interactions. Or they become dysfunctional, or they present barriers to big data practices. They need to be substituted for collective or aggregate mechanisms of control and decision making. This implies a shift from a transaction or an individual control model to a liability-model. In the transaction or individual control model the individual decides about each interaction. On the contrary in a liability-model such decision-making during operations happens more collectively or aggregated and the active role of the individual gets reduced to receiving compensation or obtaining a claim towards the big data operator.

Third, a specific method to reduce transaction costs is to move a large amount of the decision-making to the design phase and to create standardised solutions. This implies that legal or social objectives get translated into technical objectives and requirements and get integrated in the design methodologies. Standardisation is then a further collective process to lower transaction costs by creating well-understood common solutions. This reduces the needed information processing in individual decision making to deciding about which solution to choose from a set of options of which the impacts can be easily understood. Privacy-by-design has been developed above mere guidelines into a set of technical objectives with a continuously developing catalogue of threat models and solutions and into design methodologies mainstreaming these objectives alongside security objectives. Similar work needs to be done to integrate other social concerns, like discrimination.

Fourth, we are concerned with socio-technical systems and problems are not purely technical or social. Best practices are often a mixture of legal, organisational and technical elements. This implies that design objectives and methodologies and standardisation efforts have to be broadened to include the whole range of legal, organisational and technical elements.

Dealing with the political externalities also needs to take the specific impact of big data on interactions into account. On the one hand this concerns an active policy to keep markets open and competitive, where the specific ways how a company can acquire a dominant position need to be taken into account and addressed with adequate tools. On the other hand this concerns efforts by states to retain their regulative capacity. In this case adequate policies and safeguards need to be developed to preserve the balance with citizens' rights and with other states.

APPENDIX: MEASUREMENT METHODOLOGIES

The overview and evaluation of externalities makes clear that no general solutions to deal with societal externalities exist, but that these externalities are very diverse and demand specific solutions. Further, the actors involved or in the position to develop better solutions often differ as well. Solutions requiring legal change concern in the first place legislators and policy-makers. Often collective efforts are needed to develop standards or tools. On the level of the company the main responsibility consists in assessing the impact of the envisaged business model and operations, in implementing adequate safeguards to diminish the negative impacts, and in ensuring legal compliance. Capturing the benefits of big data entail an openness towards new uses of existing data holdings. This often in combination with external data sources, which necessitates an openness to new partnerships and data sharing arrangements.

Developing an extensive impact assessment and design methodology is a research project on its own and goes beyond what is possible in this task. Earlier we presented the elements of best practice already developed for such approach in the area of data protection, which needs further development and which can be extended to other problematic areas like anti-discrimination or the confidentiality of commercial information.
The agenda for such approach, based on privacy-by-design and security-by-design, contains a range of elements:
- translation of legal requirements and societal concerns into precise technical objectives
- technical, organisational and legal safeguards
- a threat catalogue identifying where and how problems to achieve the technical objectives can occur
- design patterns, presenting standard solutions for common problems or threats
- design methodologies, translating the technical objectives into threats associated with a specific project and identifying solutions with which they can be addressed.
- impact assessment methodologies, enabling the identification of risks and methods to prevent or manage them.
- standards concerning these elements

Depending on the area these elements have been developed and are in a mature stage, some initial development exists or remain to be done. In other words, a clear research agenda is present.

In this part we further present a tool to scan for problem areas. It is meant to be quick and easy in use, and does not amount to a full impact assessment. Several tools for risk and impacts assessments do already exists. Therefore we prefer not to re-invent the wheel, but rather to develop a preliminary tool which can identify areas of concern and give a quick overall view. Its results can help to decide if an impact assessment is needed and which areas need closer assessment.
This tool is inspired on several existing impact assessment and design methodologies mentioned earlier, like the PIA manuals of CNIL, the PIPEDA-tool from Canada, the LINDDUN design methodology, etc. But it does not yet amount to a full impact assessment or risk assessment. Such an assessment requires a much more detailed assessment of legal compliance and of potential threats and associated risks, which often needs sector-specific tools or input. It is therefore no substitute or framework for such assessment.

Through this self-assessment questionnaire an initial mapping can be made of potential problems, both in terms of hurdles towards benefits and negative externalities, and actions to address them.

Quickscan for externalities

Start with making an overview of data sources involved in the big data processing. Also indicate sources which would be useful, but are not accessible or for which access problems exist. Use the following question list to make a general assessment of problem areas and define under each heading the actions to take.
Optionally, you can also make a data flow diagram. Indicate for each element if problems or hurdles are present and which actions to address them are identified.

1. Potential benefits: here we look at new potential benefits and scan for hurdles

If not done yet, make an overview of useful, but not yet accessible, data sources and their potential use. Can you identify potential new uses of your data or of your data handling capacities? If yes, can you identify other data sources needed to do so and potential partners holding this data?

For each of these data sources, indicate which barriers or hurdles are present. Differentiate between the technical, semantic, organisational, legal level. Indicate which barriers are needed to limit negative impacts. Indicate which barriers are not needed for another purpose and therefore can be removed or diminished to improve interoperability.

Data source	Potential use	Barrier	Interoperability level	Purpose barrier	Actions to address barrier
Source 1		Barrier 1.1	technical		
		Barrier 1.2	organisational		
Source 2		Barrier 2.1	technical		
…		…	…		

2. The use of personal data

2.1. Is personal data used? If yes:
- List which personal data is collected or obtained. Indicate the purpose of this collection and the legal ground for this use. Indicate if the specific elements of personal data are necessary and relevant for this purpose.
- Next indicate the further uses made of this data, the purpose of this further use and the legal ground for this use. Indicate if the specific elements of personal data are necessary and relevant for this purpose. List the technical, organisational and legal safeguards implemented.
Make sure the whole life-cycle of the personal data is presented, included the deletion of the data and retention times.

- Assess who has (potentially) access to the data at the different stages. For which purpose is this access needed? How is unauthorised access or use of this personal data prevented (indicate which safeguards fulfil this role)?
- Based on these elements, make a first compatibility assessment of the further uses and the safeguards. When the further use concerns computer-assisted or automated decision-making also take the following questions into consideration.

2.2. Is there any computer-assisted or automated decision-making affecting individuals? If yes:
- Is there a review of the results by a human? Are all decisional criteria and data used transparent for this human decision-maker or reviewer?
- Indicate possible sources of discrimination on the DFD. Which auditing or control mechanisms are available at these stages? Give an overview of the safeguards to prevent any discrimination, included unintended discrimination.
- Include in the compatibility assessment also a check of the legitimacy of any automated decision-making and if the concerned persons have adequate means to object to this decision-making or to have their views taken into consideration.

2.3. Is there any transfer of personal data to countries outside the EU? Include also cloud and web-services used and assess if their use entails such transfer.
List the potential countries involved and check if the European Commission has made an adequacy decision.[185] If not, check if any of the standard contractual clauses is applicable[186] or any of the legal exceptions in article 26 (1) of directive 95/46.

2.4. Data subject rights
- Are data subjects informed of which personal data is used and for which purpose?
- Are data subjects informed of how this personal data is processed? Are data subjects informed of any automated decision-making?
- Are data subjects informed of decisional criteria and of actions taken as a consequence? Are subjects informed of the actual results and actions taken based on their data?
- Have data subjects access to their personal data? Can they correct data? Can they object to (specific) further use, in particular to automated decision-making? Can they demand deletion and how is this assessed?

3. Confidential data and use of data holdings

3.1. Make an inventory of data that is used by or open to external actors. Indicate which access to this data is desired?
- Which of this data needs to remain confidential. See next set of questions.
- Which of this data is opened in exchange for remuneration? Is protection possible by copyright and/or database rights? If not, treat as confidential data in the next set of questions and check how restricted access can be set up.

[185] See http://ec.europa.eu/justice/data-protection/international-transfers/adequacy/index_en.htm
[186] See http://ec.europa.eu/justice/data-protection/international-transfers/transfer/index_en.htm

- Is there any data to be made available as open data? Which is the envisioned purpose? How is the needed quality and interoperability to be achieved (method of publishing, license, format …)?

3.2. Make an inventory of data that needs to remain confidential. Indicate the life-cycle of the data with the level of confidentiality needed.
Assess who has (potentially) access to the data at the different stages. For which purpose is this access needed? How is unauthorised access or use of this personal data prevented (indicate which safeguards fulfil this role)? Check where data leakage can take place and through which safeguards it can be addressed.

4. The use of IPR-protected data
- Is any data in use which is protected by copyright or database rights? Indicate the different datasets and the specific IPR to which they are subjected.
- Are the uses made of the data protected by these IPR?
- If the data is protected, are there licenses acquired for the data? Is the use made of the data allowed according to these licenses?
- If not, make a list of the datasets and the rights holders. Indicate the actions to take to acquire the necessary rights.