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 D6.1:

A roadmap for big data incorporating both the research roadmap and the policy roadmap

BYTE Policy and Research Roadmap

Author(s):	Stéphane Grumbach, Aurélien Faravelon, *Inria*
	Martí Cuquet and Anna Fensel, *Universität Innsbruck*
	Scott Cunningham, *Technical University Delft*
	Rachel Finn, *Trilateral Research*

Dissemination level:	Public
Deliverable type:	Final
Version:	1.1
Submission date:	22 August 2017

Table of Contents

EXECUTIVE SUMMARY

This document presents the BYTE big data roadmap to capture the economic, social and ethical, legal and political benefits associated with the use of big data in Europe. It provides the necessary policy and research actions necessary to achieve the BYTE vision and guidelines to assist industry and scientists to address externalities in order to improve innovation and competitiveness and also incorporate the needs of civil society, NGOs and other non-profit organisations.

The BYTE project is a multi-disciplinary study of the societal impacts of big data in seven European sectors aims to define a roadmap and create a community that address and optimise these impacts.

The goal of the roadmap is to guide European policy and research efforts to develop a socially responsible big data economy, and to allow stakeholders and the big data community built around the BYTE project to identify and meet big data challenges and proceed with a shared understanding of the societal impact, positive and negative externalities and concrete problems worth investigating in future programmes.

The roadmap is the culmination of a series of case studies, analysis, expert focus groups and workshops conducted within the BYTE project. Positive and negative big data externalities in economic, social and ethical, legal and political areas have been observed and analysed in the seven initial sectors considered in the case studies (crisis informatics, culture, energy, environment, healthcare, maritime transportation and smart cities), and further extended to a total of 18 sectors via a literature review, to provide a set of risks, opportunities and recommendations of best practices. The roadmap results have been validated through two separate workshops focused on the research and the policy parts.

The policy roadmap is structured around three priorities: data governance, social good and big data infrastructure. For each priority, a set of short-term actions to perform and a set of actions to accomplish by 2030 are provided.

The research roadmap covers six areas —data management, data processing, data analysis, data protection, data visualisation and non-technical priorities— and a detailed timeframe of 5 years, which is also extended to include a mid- (up to 2025) and a long-term (up to 2030) period, to address research topics in each of these areas in order to deliver social impact, skills development and standardisation. Finally, it also identifies what sectors will be most benefited by each of the research efforts.

The present roadmap also foresees an annual deeper study of selected sectors to be taken up initially by the BYTE project partners and community members, and by the BYTE community alone after project completion. In this document, an initial analysis of the specific policy and research needs is provided for the environment, healthcare and smart city sectors selected for the first year.

1 INTRODUCTION

The Big data roadmap and cross-disciplinarY community for addressing socieTal Externalities (BYTE) project aims to assist European science and industry in capturing the positive externalities and diminishing the negative externalities associated with big data in order to gain a greater share of the big data market by 2020. In this deliverable, we present one of the primary goals of the BYTE project: a research and policy roadmap that provides incremental steps necessary to achieve the BYTE vision and guidelines to assist industry and scientists to address externalities in order to improve innovation and competitiveness and also incorporate the needs of civil society, NGOs and other non-profit organisations.

As defined early in the project, within the BYTE project we consider as a working definition that big data is that which uses big volume, big velocity, big variety data asset to extract value (insight and knowledge), and furthermore ensures veracity (quality and credibility) of the original data and the acquired information, that demand cost-effective, novel forms of data and information processing for enhanced insight, decision making, and processes control. Moreover, those demands are supported by new data models and new infrastructure services and tools which are able to procure and process data from a variety of sources and deliver data in a variety of forms to several data and information consumers and devices (Akerkar, et al. 2015, 13-14).

The work presented here is the continuation of a series of case studies, analysis, expert focus groups and workshops conducted within the BYTE project. A total of seven sectors were considered as case studies of big data practices to gain understanding of the economic, legal, social, ethical and political externalities involved in them. They comprised crisis informatics, culture, energy, environment, healthcare, maritime transportation and smart city, as presented by (Vega-Gorgojo, Donovan, et al. 2015). A horizontal analysis of the societal externalities encountered in the case studies was conducted to identify how these externalities are connected to big data practices and to each other (Lammerant, De Hert and Lasierra Beamonte, et al. 2015), to then evaluate and recommend how to address them (Lammerant, De Hert and Vega Gorgojo, et al. 2015). Based on that, the vision statement for the BYTE project was presented (Papachristos, Cunningham and Werker 2016). During the preparation of the roadmap, two more workshops have been held to obtain feedback on the roadmap draft, validate its findings and further extend the results to a broader range of stakeholders.

The roadmap, together with the community being built around it, focuses on giving good practice messages about societal issues in big data, and in particular to the environment, healthcare and smart cities sectors, which have been selected by the BYTE big data community as the ones to be addressed first. The present roadmap is expected to guide European policy and research efforts to develop a socially responsible big data economy. We also expect to contribute to the Big Data Value Association activities and priorities by bringing a societal analysis of big data impacts, and to contribute to the creation of a multidisciplinary big data community around the BYTE results that includes as well NGOs, non-profit organisations, government (and especially local government) organisations, civil society organisations and citizens.

1.1 ROADMAPPING FOR BIG DATA IN EUROPE

Elaborating roadmaps for big data is an extremely challenging task. Indeed, big data have already tremendous impact on the world in which we live. They not only redesign the way most industries are organized, but they also profoundly disrupt public administration as well as governance (Aurelien Faravelon et al. 2016a). Somehow big data services contribute to

redefine public administration to an extent which for the present might have more impact than regulators themselves in influencing big data services. Techniques used for harvesting and transforming big data and to produce new services are also evolving very fast, making it difficult to envision what technologies will be most important in 10 years. The increasing role of AI, whose usage is encouraged in the administration in the US demonstrates the tremendous changes at stake. The objective of this document is to give insights into both policy and research required to enhance the European capacities in big data.

The roadmap begins by examining global policy issues, including investments, funding and infrastructures required to take full advantage of the opportunities surrounding big data, as well as evaluating the risks for civil rights and society as a whole. We have seen the global weakness of Europe in this context in comparison with the US or even some Asian countries (Aurelien Faravelon et al. 2016a). Policy has to do with the control and potential of data, their regime, open or proprietary, their ownership, as we have shown in previous work as well as the global environment of education, research and innovation. Using the advocacy coalition framework, we were able to show the intricacy of the actions of the stakeholders, private corporations, public administrations, and citizens. The various decisions of the European court of justice illustrate very well the complexity of the work of the regulator.

Research in big data is related to a large spectrum of domains, which range from economics to computer science, from political sciences to mathematics. At the technological level, techniques are widely open, see the MEAN architecture for instance[1] (*MEAN* relies on JavaScript-based technologies — MongoDB, Express.js, AngularJS, and Node.js), but at the same time extremely complex. Data analytics, deep learning, requires extremely well trained scientists with very strong backgrounds in mathematics as well as the different technologies. The research is increasingly concentrated around emerging centers that succeed to put together researchers from different fields including the humanities (see for example, the Oxford Internet Institute). The lack of well-trained people though in this industry has been largely observed, and is a challenge for the education systems[2].

We consider successively the policy roadmap and the research roadmap, which are in large part independent. The methodologies used are largely different, although they both benefited from feedbacks from the interviews as well as from the workshops. This deliverable presents the roadmap developed by the BYTE project team in the frame work package 6 (WP6). WP6 has a twofold objective:

1. **To design a research (Task 6.1) and policy (Task 6.2) roadmap for big data** that accounts for the social impact, positive externalities and negative externalities associated with big data.
2. **To gain stakeholder consensus on the BYTE roadmap** via a workshop. The research roadmap validation workshop was collocated with the European Data Forum 2016 in Eindhoven on June 30th. The policy roadmap validation workshop was organised on September 20th in Eindhoven during the Global Forum.

The aim of the policy roadmap, as stated in BYTE's original document, is to "outline a step by step policy process for meeting the infrastructure, funding and policy needs outlined in the BYTE visions in a socially responsible way which incorporates findings related to positive and negative externalities". The policy roadmap guides the creation of framework necessary to develop a European big data ecosystem. The policy roadmap is structured around three priorities: data governance, data for social good, data infrastructure. We identify several means

[1] http://mean.io/
[2] http://www.idc.com/getdoc.jsp?containerId=249163

such as education, standardisation or data opening in order to address these three main objectives of the roadmap. After the policy roadmap validation workshop, we have identified the appropriate time horizon as 2030.

In the research part of the roadmap, we consider the positive externalities, negative externalities and social impacts associated with big data, map research and innovation topics in the areas of data management, processing, analytics, protection, visualisation, as well as non-technical topics, to the externalities they can tackle, and provide a timeframe to address these topics and prioritisation of them. We have adopted a multilayered approach that accounts for what research will develop the necessary skills, contribute to standardisation and deliver social impact to capture positive externalities and diminish negative ones in the economic, social and ethical, legal, and political areas. We have also considered how such externalities affect different sectors, and what research is more relevant to each of these sectors. The original planned time horizon for the present roadmap, 2020, has been extended to account for a detailed timeline in the upcoming 5 years after the roadmap presentation, and further mid- (2025) and long-term (2030) actions.

1.2 BYTE VISION AND SECTORS OF INTERESTS

The research roadmap aspires to contribute to the Digital Agenda for Europe for 2020 and further: what research and innovation (pillar 5), coupled with skills development (pillar 6) and standardisation (pillar 2), can maximise a positive societal impact of big data in Europe (pillars 3 and 7).

The BYTE project has identified that currently civil society organisations, NGOs and other non-profit organisations, including local administrations, are underrepresented in most big data fora. An example that is close to the BYTE project is the BDVA, which lacks representation of such stakeholders. This was confirmed in discussions at the European Data Forum 2016 with BDVA members, who showed their interest in addressing this issue and broadening their type of members. However, joining such an association as the BDVA might prove difficult for this type of organisations because of lack of time and resources. In addition, such organisations are generally reluctant to join industry-led associations like the BDVA. Therefore, the BYTE project envisions to provide BDVA with the societal analysis so far conducted, a roadmap aligned with their agenda that also incorporates the societal impacts, and further act as an intermediary between BDVA and civil society and its organisations without the need of these organisations directly joining and contributing to the BDVA.

The present roadmap and the BYTE Big Data Community thus focus on bringing good practice messages about societal issues in big data in particular sectors to industry, and will use the BDVA as one channel to achieve this.

Each year, three specific sectors will be addressed, starting with BYTE relevant case study sectors and following with sectors selected by the community. Feedback from NGOs, IGOs, academic and other civil society experts will be gathered and consolidated, and fed back to industry and policy makers. The aim is to present what are the challenges, where and what is the good practice, where are the gaps and what are the specific research and policy needs to cover these gaps. The output will be short brochures for industry and policy stakeholders within each sector.

In the first year, this will be managed as part of the BYTE project activities. After the close of the project, this activity will be taken up by the BYTE big data community, which will last at least until 2020. The sectors to focus on will be decided by key BYTE partners and the community, and we expect contributions from other BDVA members.

This will result in a direct support to industry by providing good practice information and what research and policy actions lead to a social impact. It will be specific, targeted and concise and it will follow the BYTE vision exposed in (S. W. Cunningham, Werker, and Papachristos 2016) and (S. W. Cunningham et al. 2016). The BYTE vision consists in the analysis of a set of three trends shaping the big data policy agenda for Europe: the transition, hegemony and regime of big data. These forces are applied to set of scenarios in order to identify possible futures for big data in Europe.

In the words of (S. W. Cunningham, Werker, and Papachristos 2016), "the **big data transition** involves consideration of the rapidity of technological change underlying big data", the **big data hegemony** deals with the owners of big data. For instance, will big data only fall in the hand of large companies? Eventually, the **big data regime** deals with the access to data and questions such as knowing "whether data will be a source of closed and proprietary knowledge, or whether it will be an open resource for the public good".

The BYTE vision identifies three large scale problems which the roadmap will address:

1. Given the rate of technological change in big data, **European policy setting may be partially unprepared for the positive and negative impacts resulting from a technological transition towards big data**. Should the transition be slower than expected, policy setting should do no harm.
2. Given the political economy of big data operations, **European policy setting may be poorly equipped for changes in the hegemony of big data**. If a hegemony of a few big external public or private players emerge in big data, Europe needs to exert its influence to hedge or shape the big data future. Alternatively, Europe needs to come to grip with potential futures where a diverse big data ecology is fully established.
3. Given the regime of big data operations, **European policy setting needs to be prepared to address both open and public data sources, as well as closed and proprietary protections on data**. In particular, many private European sectors are poorly prepared to transition to a potential expansion of the use of open big data.

These forces apply to a set of actors involved in policy making and influencing policy levers. (S. W. Cunningham et al. 2016) focuses on four types of actors and presents their objectives when it comes to Big data:

Small and medium legacy enterprises want to stay in business

Large enterprises want to achieve more profits than losses by developing the relevant technologies.

Policy makers want, at the same time, to benefit from big data for government action and ensure a European governance of Big data.

Consumers pursue agency, autonomy and participation.

Eventually, D 5.2 identifies four major policy levers available to public authorities:

- regulations about data openness or restrictions by governments or regulatory authorities, such as national or EU maritime authorities or NHS in the UK,
- legislation about privacy, IPR and data security issues, an aspect that seems to be central for every case we investigated.
- alignment of legislation on the national and supranational levels in sectors that go beyond national borders, e.g. the cases shipping, environment and crisis

- public investment in big data infrastructure, particularly when in the public interest such as in the case of smart cities.

The policy roadmap aims at guaranteeing the conditions to allow the BYTE vision to happen and especially the address the lack of European equipment to face the big data transition. In order to do so, the roadmap provides a set of recommendations which leverages the three forces identified in the vision in order to allow actors to pursue their goals while fostering the positive externalities of big data.

1.3 OVERVIEW OF THE ROADMAP

The roadmap consists of two parts: the policy roadmap and the research roadmap. The policy roadmap presents a set of policy actions necessary to develop a European big data ecosystem. The research roadmap identified research priorities to foster innovation. These two parts are interlinked: the policy roadmaps ensure the necessary conditions for innovation and the research roadmap identifies areas in which to invest.

The policy roadmap is structured around three priorities: data governance, social good and big data infrastructure. For each priority, we provide a set of short-term actions to performs and a set of actions to accomplish by 2030. Table 1 summarizes the policy roadmap.

The research roadmap covers six areas —data management, data processing, data analysis, data protection, data visualisation and non-technical priorities— and a detailed timeframe of 5 years, which is also extended to include a mid- (up to 2025) and long-term (up to 2030) period, to address them in order to deliver social impact, skills development and standardisation. Finally, it also identifies what sectors will be most benefited by each of the research efforts. An overview of it is presented in Figure 1.

Table 1. Summary of the policy roadmap

	Short term Actions	Actions to accomplish by 2030
Data governance	Build the foundations of international agreements on data	Balance the EU/US data relation, build a European Digital Single Market.
	Invest on the reform of data infrastructure	Turn Europe into a leader in open data and
	Start a public debate on data governance.	Build a network of civil actors addressing data governance
Social good	Promote digital literacy	Integrate data technologies in mainstream curriculums
	Assess the importance of big data on sectors such as environment and sharing economy	Help traditional business transitioning to digital business models, foster the use of big data in surveyed sectors
Data infrastructure	Promote existing data standards	Europe leading in data standards definition
	Define a strategy to improve public data capacities	EU capacities rival with other areas such as the US or China
	Invest massively on big data infrastructure and private sector	European companies are leaders in the big data sector

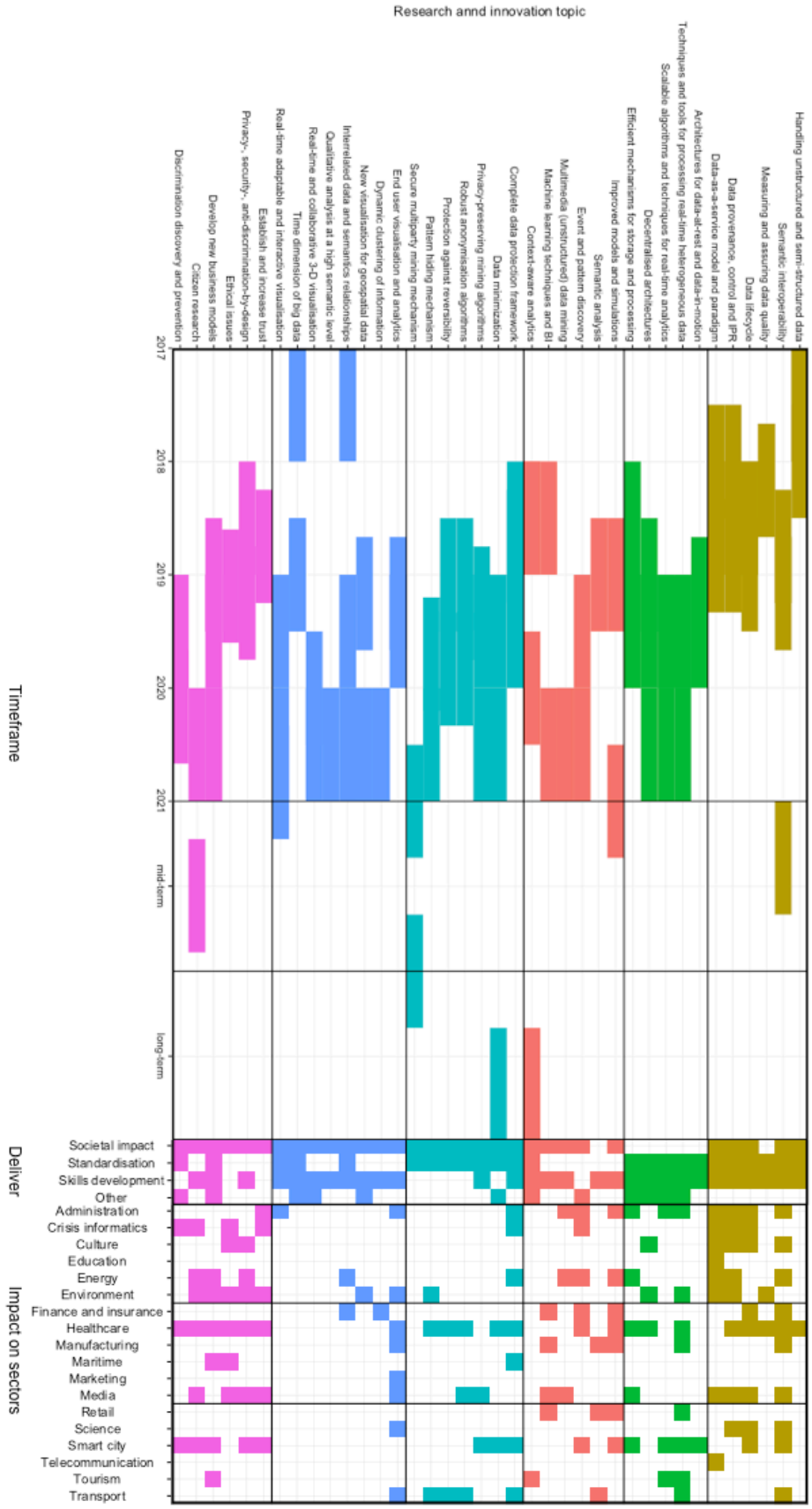


Figure 1. Overview of the research map, with a timeline to address research topics (2017 to 2021, plus mid- and long-term), how they will contribute to deliver societal impact, standardisation and skills development, and their impact to each of the sectors. Research topics are grouped in the following areas (right to left): data management, data processing, data analysis, data protection, data visualisation, non-technical priorities.

1.4 ORGANISATION OF THE DOCUMENT

The rest of this document is organised as follows. The policy roadmap is presented first followed by the research roadmap. Both roadmaps begin with an overview of their scope and methodology, followed by the contextualisation of the roadmap within BYTE and the wider policy space and research priorities in Europe. Each then outlines the key policy and research priorities for Europe, with both short term and medium term goals.

2 A BIG DATA POLICY ROADMAP FOR EUROPE

2.1 BACKGROUND

2.1.1 Current European context and goals for a policy roadmap

Europe is dependent on foreign digital companies. Policies, such as the "right to be forgotten" are under elaboration. It is not clear yet if they empower Europe or foreign companies or countries. In some European countries, such as France or Spain, the discussion around "digital sovereignty" is vivid and answers are yet to come (Zeno-Zencovich 2016).

On the other hand, Europe hosts a massive wave of investments in the big data sector and a lively discussion on the right decisions to make (EU 2016). The Big Data Value Association (BDVA), for instance, provides five priorities to tackle in order to sustain the development of big data in Europe[3]. Namely:

> "**Data Management**": foster data interoperability and availability
>
> "**Data Processing Architectures**": build architectures adapted to new technological constraints (flows of data, heterogeneity), etc.
>
> "**Data Analytics**": Upgrade existing analytics to build, for instance, "prescriptive and predictive analytics", improve data models, etc.
>
> "**Data Protection**": Define better protection models and mechanisms, such as data encryption.
>
> "**Data Visualisation and User Interaction**": build reusable components for data visualisation.

These priorities are in line with objectives defined at the European Commission level. For instance, the European Innovation Partnership on Smart Cities and Communities, which brings together European cities, leading industries and civil society members to build smarter cities, identifies data architectures and analytics as key tools[4]. If the BDVA does not explicitly mention data policy, it remains a central lever to shape the future of big data and incentivise the development of a big data ecosystem in Europe[5].

Yet, each European country is still struggling with the definition of a data policy. For instance, in Spain, the famous opposition to Google News led to the shutting down of the service which remains available everywhere else in Europe. On the contrary, in the UK, an agreement between the NHS and Google allowed the company to apply its Artificial Intelligence software - Google DeepMind - to health data[6].

The *Spanish Copyright Act* exemplifies that regulation, in the field of big data, can be harmful to economic activities even if the consequences of the Spanish regulation are still debated. Indeed, some authors highlight that charging aggregators that quote news providers is counter-productive to the extent that news providers benefit from the traffic to the aggregators. Well-known newspapers such as *El País* have refused to charge news aggregators. At any rate, the Spanish regulation deprives Spanish news providers from a major aggregator and forbids the development of news aggregators in the country. It also highlights the conflicting values and interests between, on the one hand, digital companies which operate worldwide and offer free

[3] http://www.bdva.eu/sites/default/files/EuropeanBigDataValuePartnership_SRIA__v2.pdf
[4] https://ec.europa.eu/digital-single-market/en/smart-cities
[5] http://dataforpolicy.org/
[6] https://www.newscientist.com/article/2086454-revealed-google-ai-has-access-to-huge-haul-of-nhs-patient-data/

services and, on the other hand, nation states which have to rethink their tax collection procedures and are unable to impose their values to private companies.

In contrast, the agreement between Google and the NHS relies on the hopes provided by artificial intelligence. Google DeepMind's predictions are expected to provide insights on diseases and fuel research. Yet, this agreement is deeply criticized as it could reinforce Google's almost monopolistic position in data processing. It also provides a private company with a huge influence on public health and public good and showcases that public services are dependent on private organisations to perform their mission.

More broadly, the discussions around a European digital single market exemplify the raising awareness of the need to suppress the digital barriers between member states[7]. A single digital market is especially important to allow Europeans to travel freely through Europe - without paying roaming fees for instance - and benefit from a uniform level of data protection over the European territory while accessing the same digital content all over Europe.

Nonetheless, data economy is a global phenomenon which calls for global agreements. Most date companies are multinationals and some technological frameworks - such as cloud computing - are inherently cross-border: cloud computing implies the ability to access one's data without taking care of one's location or device. Big data also is a matter of international law. Data processing is a global activity. Yet, the globalization of data flows clashes with national - and thus diverse - legislations and practices on data. Negotiating agreements between several areas of the world and between states and companies is thus necessary to allow big data as a global business. The so-called "Safe harbour principles" exemplify such an agreement. Designed in Europe in the early 2000's, the seven principles (namely notice, choice, onward transfer, security, data integrity, access, and enforcement) were meant to describe the adequate protection of data for European and Swiss citizens. These principles also appeared as business facilitators: the European Commission stated in 2000 that American companies respecting these principles were allowed to process the data of European and Swiss citizens when the EU considers that the US do not provide an adequate level of data protection[8].

The Safe Harbor appears as a pragmatic agreement between legal and technical experts which allows the development of a transatlantic digital economy. Yet, a recent court judgement elevated the location of customer data from an issue which was addressed in back rooms by policy specialists to a topic needing urgent redress by law. Indeed, in 2015, the European Court of Justice stated that the Safe Harbour Decision was invalid after an Austrian citizen complained that his Facebook data were not adequately protected. Snowden revealed the cooperation between the US intelligence services and digital companies under the Patriot Act[9]. Several audits also established that self-certification was a weak mean to ensure that companies actually complied with the Safe Harbor principles as most companies did not actually implement the principles. Several discussions between the US and the EU led to the "EU-US Privacy Shield" which was adopted in 2016 and was necessary to keep on businesses involving European data[10]. This Privacy Shield still relies on self-certification by American companies.

Eventually, beyond international agreements, the development of data economy causes several policy challenges. First, the relation between public and private actors when it comes to sharing data needs to be clarified. For instance, ride hailing companies such as Uber possess data necessary to better govern cities by solving traffic jams for instance. Yet, if some of them

[7] https://ec.europa.eu/priorities/digital-single-market_en
[8] http://2016.export.gov/safeharbor/
[9] http://www.bbc.com/news/world-us-canada-23123964
[10] https://www.privacyshield.gov

collaborate with local governments, mot local governments conflict with transport applications[11]. When it comes to data protection, companies such as Apple or WhatsApp define new standards of encryption and governments of intelligence services have to cope with them[12].

Eventually, the data economy offers opportunities to citizens, for instance to earn more money or connect with users and services. The policy framework of the so-called sharing economy is still under elaboration and should be refined to help citizens develop their activities while instituting fair rul.

2.1.2 Goals of a policy roadmap

The European policy on big data is still under elaboration. Clearly, a policy roadmap should minimize the negative externalities of a big data policy. Such a policy could forbid the development of innovative technologies and prevent the digital development of Europe or submit it to a more intense influence of foreign companies. It could also foster the development of a big data ecosystem that could benefit to private and public actors and citizens.

The policy roadmap focuses on safeguarding the conditions for the creation of a European big data infrastructure while promoting social good and ensuring a fair data governance.

2.1.3 Existing roadmaps and BYTE approach

Corporations and public actors are all involved in designing big data roadmaps to make the most out of the data revolution. Google Search retrieves more than 10,000 results for the keywords “Big data roadmap” and “Big data whitepaper”. We select a handful of roadmaps which illustrates the directions under consideration. We also select the roadmaps according to their authors in order to reflect the preoccupation of public bodies and policymakers, academics and private actors.

Table 2 presents the main roadmaps we have surveyed and, for each roadmap its main focus and if it includes policy-oriented recommendations. Policy is only directly addressed by 2 roadmaps, even if two other roadmaps mention policy-related issues, such as standardisation and explicitly state the need for a policy roadmap. When policy is addressed, authors are unanimous about the need to build the conditions of a big data ecosystem, such as standardisation, and values, such as privacy. Opening data in order to avoid the creation of data silos and open the existing ones is also an acclaimed priority.

The policy roadmap aims at answering the need for a policy roadmap expressed in the literature. It benefits from existing roadmaps as it adapts existing recommendations - for instance related

[11] https://www.boston.com/news/business/2016/06/28/uber-data-boston-wants
[12] http://www.cnbc.com/2016/03/29/apple-vs-fbi-all-you-need-to-know.html

Table 2. Existing roadmaps.

	Authors and year	Main Recommendations	Policy-related recommendation
Big Data: Seizing Opportunities, preserving values	Executive office of the President (US), 2014	Policy Framework for big data	Preserving privacy-related values, leverage issues such as discrimination and use data as public resource.
Innovation Roadmap	BDVA, 2016	Technical priorities (for instance: data visualisation or data management).	No direct policy-related recommendations, even if standardisation, for instance, is addressed.
Big Data Analytics – Roadmap, Energy Sector	Think Big Analytics, 2013	Staff training, Data management	None
Roadmap to extract value and knowledge from medical data	Association of the British Pharmaceutical Industry, 2013	Create a sustainable data econsystem, for instance by building capability and capacities.	No direct policy-related recommendations, even if standardisation, for instance, is addressed.
Big Data and Internet of Things: A Roadmap for Smart Environments	Nik Bessis, Ciprian Dobre, 2014	Technical priorities (Data models, techniques and applications)	None
Roadmap for Big Data Research	Big Project, 2015	Cross-sectorial Requirements Analysis for Big Data Research	Open data, European Digital Single Market, Education
Enterprise big data predictions	Oracle, 2016	Foster the development of data-intensive services	None
The Big Data Payoff: Turning Big Data into Business Value	Informatica, Cap Gemini, 2016	Roadmap to achieve data-oriented business objectives	None

to privacy management - to the current European context. It also relies on the knowledge built by the BYTE project: the policy roadmap explicitly tackles the issues identified in the BYTE vision (S. W. Cunningham, Werker, and Papachristos 2016) and BYTE case studies (Vega-Gorgojo, et al. 2015). For instance, the case studies identified that sectors such as crisis informatics - dealing with the use of social media in crisis management - and the environment sectors may be ill prepared to face to big data transition. They also underline the need to improve digital literacy (Lammerant, De Hert and Vega Gorgojo, et al. 2015).

2.2 BUILDING THE ROADMAP

The policy roadmap for big data in Europe focuses on securing the necessary conditions for the development of a European big data ecosystem. We have built the roadmap in two phases. First, we have drafted the roadmap and then we have submitted it to a group of experts in order to refine and validate it.

2.2.1 Methodology

Our methodology is inspired by the current trend of technology roadmapping (Wimmer, Cristiano, and Xiaofeng, n.d.) (Moehrle, Isenmann, and Phaal 2013; Bernal et al. 2009). The EGOVRTD2020 project, for instance, provides insights in technology roadmapping for building e-governments[13]. (Wimmer, Cristiano, and Xiaofeng, n.d.) especially insist on the necessity for a policy roadmap to address long-term issues and connect these issues with concrete technological and policy measures. In policy roadmapping, expert knowledge is widely acclaimed as a precious resource collected with tools such as workshops or interviews.

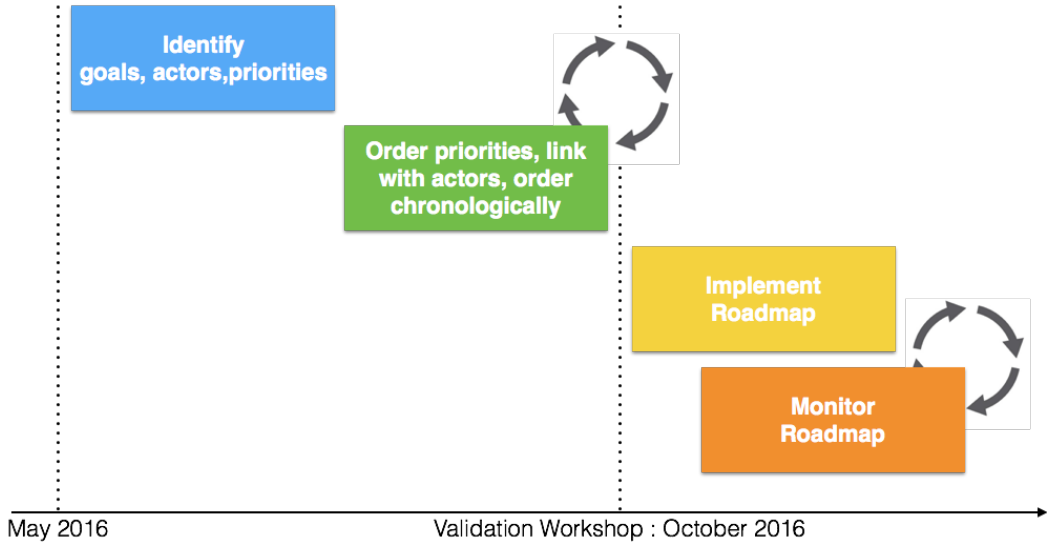


Figure 2. Policy roadmap methodology.

Figure 2 presents our overall methodology, divided in four steps. At first, we identify **Goals, actors and priorities.** This step aims at setting out the objectives the roadmap will help achieving, their priorities and the relevant actors. This step relies on a mix of the application of the Advocacy Coalition Framework (Sabatier 1998) to a set of case studies, meetings and workshops with experts. For instance, we have met up with local policy makers and data officers from the city of Lyon and taken part to workshops organised by the city to design a digital and innovation plan. This step helped us to identify the actors responsible for designing and implementing a data policy and draft a first step of priorities.

Then, we ordered the priorities according to their importance and chronologically, we also assigned them to a set of actors. This step aimed at ordering the objectives and their means identified at step 1. This step also defined indicators or desired outcomes necessary to monitor the roadmap. This step relied on expert judgement, meetings and workshops. We

[13] https://www.uni-koblenz-landau.de/en/campus-koblenz/fb4/iwvi/agvinf/projects/completedprojects/egovrtd2020

refined our draft and designed an online survey[14] to test the relevance of the priorities we identified and their time horizon. The survey was sent to the BYTE mailing list and to a set of selected experts in law, computer science and privacy. We also held a role-playing game during the 2016 Global Forum in order to get insights from a set of experts and validate our draft.

The roadmap will live through two more steps. First, the **roadmap will need to be implemented.** This step belongs to the actors who will put into action the recommendations formulated in the two previous steps. The Byte project members and the BYTE Big Data Community will also be responsible for ensuring that the roadmap is implemented and monitoring this implementation. Monitoring the roadmap relies on the indicators defined to assess the roadmap's recommendations and make the necessary adjustments. Monitoring the roadmap should help making the roadmap evolve and adapting it to the evolution of big data practice in Europe.

2.2.2 Actors

Our list of actors involved in policy-making or influenced by policy derives directly from D 5.1 and D.5.2. We have extended this list and deepened our understanding by applying the Advocacy Coalition Framework[15] (Sabatier 1998) to a range of big data situations such as the Spain vs. Google case, the Spanish Copyright Act and examples drawn from smart cities (S. Cunningham, Faravelon, and Grumbach 2016).

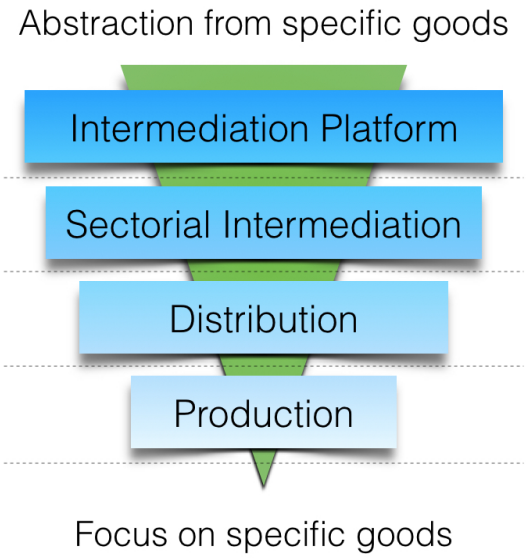


Figure 3. Classification of actors acording to their level of abstraction.

We especially refine the list of actors presented in D 5.1 when it comes to enterprises. Indeed, in (Faravelon et al. 2016b), we show that, from a data perspective, companies range from a low level of abstraction to a very high one. Industries fully involved in constructing and distributing their goods - such as the press - are ranked as low abstract businesses. On the contrary, digital platforms, which mainly offer an infrastructure on top of which other service providers may develop other services are defined as abstract businesses. We call such platforms

[14] https://aurelienf.typeform.com/to/x3MHUW

[15] The advocacy coalition framework (ACF) allows to model the interactions between a set of actors in a policymaking environment. A "coalition" refers to a group of actors who share the same beliefs and try to translate them into actions through policymaking. In an environment, several coalitions may compete.

"intermediation platforms" as most of their activities consist in intermediating between users, services and service providers.

Intermediation platforms are central in the current data economy and our aim is to understand why they are so. To do so, we build a model of intermediation activities. We propose to distinguish between different types of intermediation according to their relations to the products or services they deal with. These relations strongly influence a platform's relations with its users.

An "abstract" service does not focus on a specific usage. For instance, a social network is an abstract service as it offers a wide set of functionalities not restricted to an economic sector, and it allows to build other functionalities on top of it using its API. In contrast, an online shop provides a limited number of functionalities. We distinguish four levels of intermediation activities, ranked according to their degree of abstraction in ascending order (Figure 3). Levels can overlap: a company may develop services in several categories of activities.

At the bottom of the hierarchy, the "production level" encompasses industries that produce goods or services and sell them online to their customers, essentially with a very restricted form of intermediation between their services and their customers. The press constitutes a good example of such industries, with a direct relationship to their readers.

One level above, distribution corporations commercialise goods produced by others. Netflix is an example of that level, giving mostly access to cultural products it does not produce but distributes to its customers.

The next level, sectorial intermediation, includes corporations which provide online services which allow their users to connect with specific goods or services. The search engine belongs to that category together with actors, such as Blogspot or LinkedIn and more generally online dating sites or job sites.

Eventually, the highest level, intermediation platform, is constituted by corporations that offer an ecosystem on top of which others can build and distribute their services. Facebook and Google, for instance, are the most prominent platforms. At this level, corporations offer a sort of global operating system disconnected from physical supports, that allows the development of unbounded types of activities.

Categories may overlap. Amazon, for instance, is mostly in the distribution category but operates as well as sectorial intermediation, while Netflix mostly distributes digital contents but also produces some. For production systems, the intermediation activity might be shallow. Transport operators can be seen as intermediating between drivers and clients for instance. Yet what is of interest, is that although the intermediation might be shallow, there is a possibility of "dis-intermediation" of the activity, which is what is going on in the transport sector in particular. Companies, may challenge traditional transport business models - which focus on production and distribution activities - by performing intermediation. Carpooling platforms or ride hailing applications are examples of such challengers.

To us, legacy businesses which do not collect or process data and data companies have very different attitude towards big data. Data companies, such as intermediation platforms, are naturally immersed in the data economy and base their services on data collecting and processing. Legacy businesses have to cope with the big data transition and may either try to adapt to this transition or fight its effect. The policy roadmap fully acknowledged the disruptive effects of data economy and fosters innovation while ensuring fair rules of the game and leverage potential imbalances.

2.2.3 Legacy businesses

Existing enterprises from the pre-digital economy face major challenges for their business models. They lobby and push to protect their models and their markets. They face major loss and their assets may lose their value. Taxis are good examples of traditional actors. Everywhere, ride hailing applications challenge taxis. Drivers are losing their jobs and the value of assets such taxi medallions is called into questions. In reaction to this challenge, taxis protest against new actors and hope that policymakers will help them to protect their income.

For legacy businesses, the challenge is to survive the digital disruption and to transition to a digital business model. Some companies, such as Michelin, are successfully undergoing such a transition[16]: the company use the data it possesses on drivers and territories to provide new services. On the contrary, some businesses, such as low quality hotels are forced to shut down because of the rise of digital services (Byers et al., n.d.).

2.2.4 Digital businesses

Digital enterprises, such as start-ups, challenge existing rules. They push new vision of the customer-producer relationship, economic models - such as the "sharing economy" and relation to institutions (Orsi 2012). Companies, such as Uber or Airbnb, strongly promote freelance work, for instance, and challenge fields such as taxation or trade unions[17]. In a way, data companies may be considered as empowering users with new economic means. They may also be seen as promoting models of work relations without social protection and as opponents to policy makers.

2.2.5 Fostering positive externalities of big data policies

Negative and positive externalities can be seen as two faces of the same coins. For instance, Airbnb empowers end-users who can make money from their houses but challenges some hotels (Byers et al., n.d.). The company shakes local rules but it is also a first-class partner when it comes to collecting local taxes[18]. Uber is well-known for offering an efficient and user-friendly services but it forces taxi drivers to redefine their jobs and challenge their income.

In many sectors - such as artificial intelligence or health - big data are expected to bring smarter and more efficient services. Yet, the seemingly transparence of personal data may arouse fears about privacy management for instance. From a political point of view, powerful big data companies can be both considered as partners - in Boston, the town services collaborate with Uber, for instance - or opponents. The opposition between Apple and the US government on encryption standards, for instance, is well-known. The policy roadmap addresses these issues to encourage the positive externalities of big data.

[16] http://www.michelin.com/fre/innovation/recherche-et-developpement/michelin-et-innovation-en-chiffres
[17] https://www.theguardian.com/technology/2014/dec/03/amazon-mechanical-turk-workers-protest-jeff-bezos
[18] https://www.airbnb.fr/help/article/1383/h-bergement-responsable-en-france

2.2.6 Drafting and validating the roadmap

The policy roadmap was validated according to a two steps process. First, an online survey helped validating a preliminary set of priorities and identify the relevant time horizon for the policy roadmap. In September 2016, the survey was submitted to a group of 6 experts in law, privacy and computer science in September 2016. It was also sent to the BYTE mailing list[19].

We collected 14 answers. Respondents belonged either to academia or the private sector. Figure 4 presents the field of expertise of respondents. Figure 5 presents the countries of origin or the respondents.

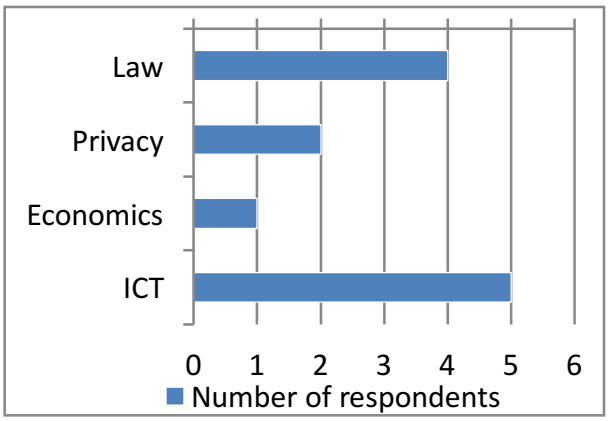


Figure 4. Field of expertise of the respondents.

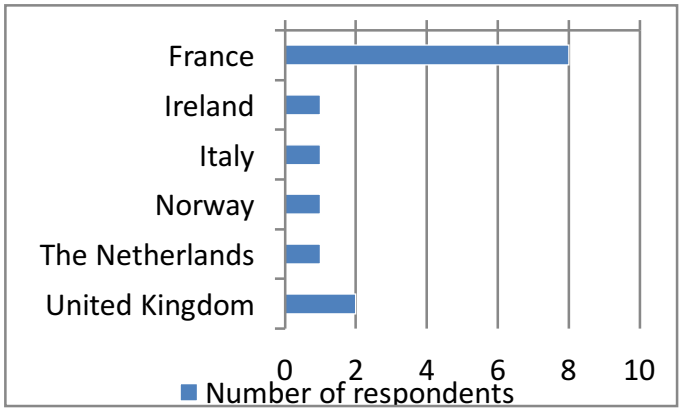


Figure 5. Origin of the respondents.

The answers to the survey confirmed the relevance of the three challenges we identified in policymaking for big data, namely data governance, business development and social good: all respondents rated them as high-level priorities. Respondents suggested refinements of these priorities. Privacy management and data access are the most cited refinements.

The survey helped us identify the relevant policy tools, externalities and actors responsible for implementing a big data policy.

As shown on Figure 6, all the respondents cite "standards and norms" as the main policy tools to use. Examples of relevant norms and standards include data exchange formats.

Respondents rank "business development" as the first and most important externality.

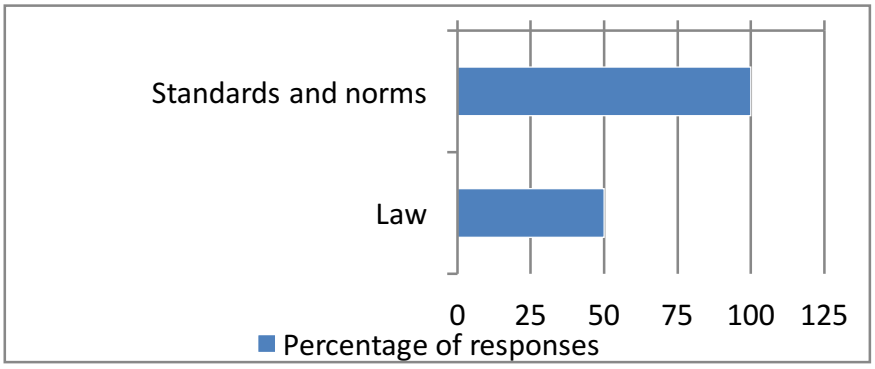


Figure 6. Policy tools.

[19] The survey is available in the Appendix of this document.

Respondents state that a big data policy should create the conditions to build a powerful European infrastructure for big data.

Eventually, respondents were asked to identify who should be responsible for addressing data governance, business development and social good. Respondents show a high level of expectations towards policymakers and acknowledge the prominent role of the private sector. Eventually, the respondents identify 2030 as a relevant time horizon for the policy roadmap. Respondents refer to some actions that are already happening –for instance, discussions of transatlantic agreements– and acknowledge the need for long-term actions to build a European digital single market, for instance.

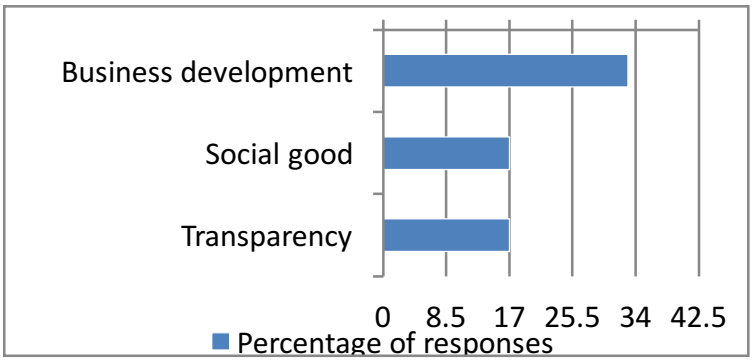


Figure 7. Externalities of big data.

The answers to the survey helped us draft a roadmap made of a set of preliminary objectives and policy actions. This draft was validated during a dedicated workshop organised in conjunction with the Global Forum in Eindhoven on September 20th. During the workshop, 23 participants from backgrounds ranging from industry to academia and from several European member States, were asked to play a role game. Roles were defined according to the list of actors we have presented. A set of objectives, deriving from the vision, was assigned to each role and participants were asked to find potential partners to help them achieve their goals.

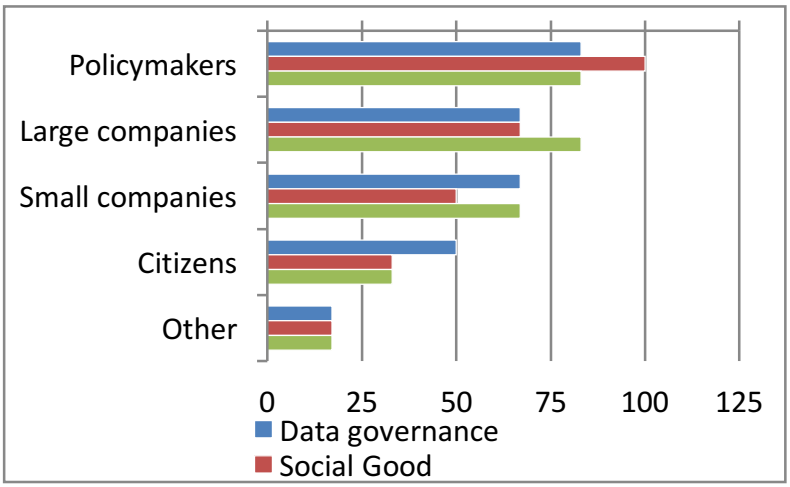


Figure 8. Actors responsible for implementing a big data policy.

Specifically, participants were asked:

- To embody a role with objectives to achieve
- Describe the ideal situation to make the most out of the big data transition while achieving their objectives and finding partners. This situation is called the "target situation".
- Identify breaks to this ideal situation
- Identify solutions, collaboration points and trade-offs to make this situation happen.

Participants were asked to log all their actions on post-its and move them close to their partner's. These logs allowed to obtain the description of a collective ideal situation and the breaks which prevent from implementing it.

Figure 9 presents the results of the role game: on the left-hand side, we see the "target situation": actors placed their cards close to potential collaborators and formed coalitions. For instance, Taxi placed themselves close to Hotels. Their coalition aims at business development.

On the right-hand side, participants were asked to identify policy-related issues and their potential solutions. For instance, the actor "privacy-aware citizen" identified the lack of understandable data format and exchange standards and the lack of certification as two obstacles to privacy management. The actor makes four propositions to improve the situation:

- Design a new regulation as soon as possible
- Develop standards with regulators and NGOs in the next year
- Develop standard and frameworks with companies such as Apple or Samsung in the next year
- Create grassroots political parties in the next ten years.

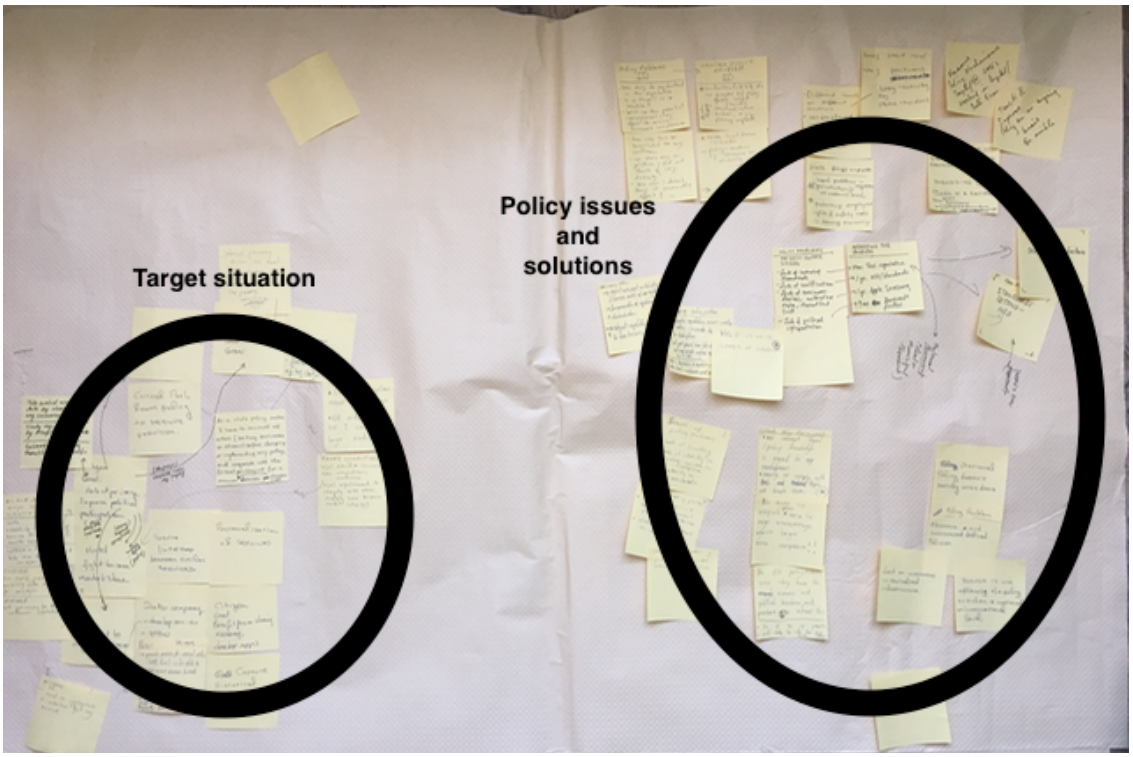


Figure 9. Traces from the workshop.

The main result of the role game is a description of an ideal European policy situation and a set of obstacles to its implementation and potential solutions. In this idea situation, the autonomy of actors, such as the citizens, is encouraged. So is their ability to claim for the respect or right such as privacy. The participants to the workshop were unanimous about the need for a EU wide data policy that would state clear rules and allow control over data. This policy would rely on data standards and exchange formats and foster the development of a European infrastructure. Participants also all agreed about the need to foster social good through the development of big data. Privacy management, resource management and a fair distribution of the benefits of the data economy were unanimously cited as components of social good.

Eventually, participants expressed a high level of expectations towards policymakers and regulators. They also underlined that some actions should be conducted as soon as possible - such as helping traditional businesses transitioning to digital business models - but some deep changes - such as the reform of citizen representation - need a long-term horizon.

2.3 THREE PRIORITIES FOR BIG DATA IN EUROPE

The policy roadmap aims at safeguarding the conditions to implement the ideal situation identified during the validation workshop. We identify three priorities to address in order to implement these situation: (1) data governance, (2) Emergence of a powerful European infrastructure for big data, (3) big data for social good. We chose 2030 as the time horizon of the roadmap according to the results of the survey and the validation workshop.

We now detail each priority and its associated set of actions and recommendations. For each priority, we draw a short background, the main challenges to address, and the outcomes we could expect which provide a set of indicators to monitor the implementation of the roadmap.

2.3.1 Data governance

Background

Data governance refers to data management and the monitoring of this management (Ladley 2012). It addresses matters such as access to personal data and privacy management or the access to corporate data and open data. Data governance implies shared decisions about the management of data. As such, it involves public partners, private partners and civil society. The focus on data governance extends the findings of the evaluation and the horizontal analysis of big data's externalities (Lammerant, De Hert et. al, 2015, Lammerant, De Hert and Vega Gorgojo, et al. 2015). Indeed, the horizontal analysis shows that privacy should be addressed to leverage the externalities of big data. The evaluation of these externalities also concludes that new legal means, for instance to protect privacy, should not hinder business opportunities.

Currently, data governance often derives from backdoor negotiations or conflicts. For instance, Snowden's revelations on the PRISM surveillance program and several examples such as "Europe vs. Facebook" show that privacy matters to citizens[20]. Cases such as *Google vs. Spain* or the Spanish Copyright act show that possession and access to data is the source of a strong power[21]. Both cases, which involve Google, show that a search engine - which possesses access to its users queries and may provide "personalized answers" is able to shape the access to data and search results and influence the economic activity of other actors - such as online newspaper.

Data governance tackles three questions:

- **Who owns a dataset?**
- **Who has access to a dataset?**
- **What can one do with a dataset?**

Data governance also relies on adequate monitoring procedures.

Currently, data governance is a conflicting field. The opposition between Apple and the US government on matters such as encryption, for instance, show that companies and public bodies are yet to find agreements on data formats and sharing[22]. Data governance is especially important as it determines data sharing and processing possibilities in a world where some corporate datasets may be important for social good. For instance, processing search queries may help to understand and protect public health (Pollett et al. 2016). The wealth of data possessed by public actors and private companies on people determines their possibilities in terms of privacy protection.

[20] http://europe-v-facebook.org
[21] http://curia.europa.eu/juris/liste.jsf?num=C-131/12
[22] https://theintercept.com/2016/02/23/new-court-filing-reveals-apple-faces-12-other-requests-to-break-into-locked-iphones/

Data governance also leads to reforming existing practices and procedures. For instance, organisations - be them public or private - often possess dataset "in silos". Opening these silos may be an important challenge to help make the most out of data (Stott 2014).

Challenges

From the literature review and the validation workshop, we identify four main topics which should be addressed to govern data while fostering the development of a European data ecosystem:

- **Facilitating intra-European data transfers**: doing so requires the harmonisation of the European digital market and data protection policies across member states.
- **Facilitating extra-European data transfers**: doing so requires signing agreement with other countries, especially to guarantee an adequate level of data protection.
- **Foster innovation:** innovation relies on the right investments and the adequate policy framework to allow data exchanges and transfer.
- **Fostering transparency**: this challenge addresses the possibility to use data in order to improve the transparency of governmental actions for instance or make data policies clearer.

Recommendations

For each challenge, we provide a set of recommendations in order to strengthen the European big data ecosystem. Table 3 provides an overview of these recommendations, which are intertwined. The development of a privacy framework - which involves agreement on the notions such as the "right to be forgotten" and an inquiry into the definition of personal data, for instance, is necessary to gain the trust of citizens and empower them. Developing such a framework requires a level of transparency allowing citizens to understand their possibilities in terms of privacy protection. It also necessitates a public debate on data governance and education of users. Yet, developing such a framework also relies on the collaboration of public and private actors. There are examples of foundations of such frameworks such as Google's form for the enforcement of the "right to be forgotten". Such frameworks could help citizens know what companies and public bodies know about them and manage access to their data.

At any rate, international agreements on data sharing are necessary as the data economy is a global phenomenon. Agreements between Member States should be signed as soon as possible to provide a uniform European data landscape. From an international perspective, the conflicts around the Safe harbour and the more recent Privacy shield, show that better agreements are necessary. Such agreements are necessary as European is strongly dependent on foreign data companies. Safeguarding the protection of the data of European citizens is thus crucial. Yet, fostering innovation, and the development of European data companies, remains a priority as such companies could be first-class partners for European policymakers.

As a result, the aforementioned challenges call for the development of public/private partnerships. The development of services such as Google Flu or Facebook's safety check show that, as data companies are in direct contact with their users, they collect very large datasets and gain the ability to develop services for the community. If public bodies want to be able to develop, use or benefit from these services, they must partner with data companies, or, if they want to bypass them, turn themselves into data actors. As a result, the data governance priority strongly relies on the development of a relevant data infrastructure in Europe which would promote the use of data for social good.

Table 3. Policy actions for data governance.

	Short term actions	Actions to accomplish by 2030
Facilitating intra-European data transfers	Build foundations of a EU single market	Sign an Agreement on a EU single market
Facilitating extra-European data transfers	Design a EU/US agreement on data	Balance the EU/US data relation
Foster innovation	Invest in government as a platform	Develop everyday services for citizens
	Start large-scale private/public discussions on data management	Balance the private/public data relation
	Foster open data and abolish data silos	Turn Europe in a leader in open data
Fostering transparency	Start a public debate on data governance	Build a network of civil actors addressing data governance
	Raise awareness about privacy management and tools	Turn Europe in a leader in privacy-enhancing technologies

2.3.2 Social good and citizen implication

Background

The data transition has already disrupted - or promises to do so - a large wealth of sectors. Participatory sensing and data mining are two techniques proved useful to help urban planning. In health, private services, such as Google Search, have demonstrated that data seemingly unrelated to health - such as search queries - could help predict the onset of an epidemic[23]. Data processing is an essential part of genetic research and national health systems are tempted to partner with data companies to process their data, at the risk of facing privacy concerns[24]. From

[23] https://www.google.org/flutrends/about/
[24] https://www.england.nhs.uk/2016/01/embracing-innovation/

an economic point of view, new paradigms, such as the so-called sharing economy, challenge existing modes of profit repartition and traditional business models.

Eventually, two challenges have drawn our attention through our work. First, the human race faces global issues, such as global warming, for which data are considered as a useful tool to design new solutions (Mayer-Schönberger and Cukier 2013). Global platforms could help monitor resources consumption and pollution as some of them are directly involved in the transportation sector for instance. They could also provide useful data processing mechanisms - the development of artificial intelligence, for instance is expected to "make the world a better place" (Knight 2016) - and means to reach people globally and address them incentives.

Eventually, most western democracies face a drastic drop of citizen participation in elections for instance. The data transition could be the opportunity to develop innovative ways to interact with citizens and foster their participation. Electronic voting, for instance, is a much-debated topic (Kersting and Baldersheim 2004). Yet, developing online services and turning government into platforms could also be a way to help citizens interact more easily with institutions (O'Reilly 2011). So could the institution of representatives from the civil society in charges of discussing data-related issues.

Table 4. Policy actions for social good.

	Short term actions	Actions to accomplish by 2030
Citizen participation	Promote digital literacy	Integrate data technologies in mainstream curriculums
	Audit current security procedures	Implement reliable and transparent security procedures
	Assess the benefits and the risks of sharing economy	Europe helps legacy business adapt to the uberization and adapts its social system and taxation system
Resources management	Assess the importance of big data on urban planning	Diminish digital divide and social divides in smart cities
	Assess the importance of big data on resource management	Make Europe lead a global platform for climate change management
Sharing the benefits of the data transition	Assess the benefits and the risks of sharing economy	Europe helps legacy business adapt to the uberization and adapts its social system and taxation system

Challenges

We identify four challenges to address in order to improve social good thanks to the data transition. These challenges all belong to the social benefits of big data identified throughout Byte's case studies (Cuquet et al. 2016).

- **Security**, which is a trade-off between values and procedures such as freedom and surveillance for instance.
- **Resource management** relying on tools such as participatory sensing or global platforms could help monitor resource consumption, identify potential areas of economy and design new solutions to water shortages or pollution (Chalh et al. 2015). Monitoring traffic, incentivizing green transportation modes could help improve life quality. Real estate management could be rationalized to help landlord valorise their assets and help potential tenants, for instance, find places to rent. Eventually, social divides could be monitored and leveraged through urban planning.
- **Citizen participation.** The data transition could help design new mode of citizen participation beyond elections. Fact-checking or participatory sensing are but two examples of participation modes.
- **Sharing the benefits of the data transition**. The data transition brings new business models that require rethinking the way we share the benefits of traditional activities and data-based ones.

These challenges derive from the externalities identified in BYTE case studies. Specifically, these challenges build on the analysis of crisis informatics, environment and smart city. The case studies show, for instance, that social media may be a useful tool to handle crisis, even though they raise issues related to privacy for instance (Cuquet et al. 2016).

Recommendations

Table 4 provides an overview of the recommendations we draw from our work in order to address the four aforementioned challenges. Surveys show that citizens are still ill-informed on matters such as privacy mechanisms and that digital literacy could be improved (Centre for the Advancement of Social Sciences Research HK Baptist University 2013). Addressing these points is necessary to allow citizens to benefit from the data transition. This could help to turn citizens from mere users of technologies into actual actors. Providing a safe space also is necessary to allow citizens to live a peaceful life and the data transition could help doing so (Chen, H., Chiang, R. H., & Storey, V. C. 2012).

Sharing economy is both a highly-praised paradigm and a feared one. It both enables citizens to monetize their assets and submits them to new and potentially harmful work relationships. It disrupts existing business models and provides new economic opportunities. Yet, taking advantage of these opportunities require fair rules and adequate means for workers' protection (Schor 2014).

Eventually, resource management is a global issue that call for a global solution. The analysis of the environment sector during BYTE's cases studies show the effectiveness of data-intensive techniques (Cuquet et al. 2016). Indeed, techniques such as data processing or targeted incentives could help model and influence the current situation. For instance, data mining is praised as a tool to model and understand climate change (Hampton et al. 2013). From a hardware perspective, so-called green computing, which limits the consumption of resources by data activities, probably is a fruitful way to explore too (Vikram and Shweta 2015).

Quite obviously, this priority relies on a relevant infrastructure to make the most out of the data transition. We now turn to the development of this infrastructure.

2.3.3 Emergence of a powerful European infrastructure for big data

Background

Europe is strongly dependent on foreign systems when it comes to data of the Internet. For instance, in France, the top 25 web platforms - i.e. the ones which attract most users - are all American (Faravelon et al. 2016b). As a result, Europe as a whole and specific member states often conflict with platforms on matters such as tax collection or values.

Nonetheless, platforms' large population of users and, consequently, large datasets allow them to gain some governance power (Faravelon and Grumbach 2016). For instance, platforms are able to influence car traffic and reshape neighbourhoods[25] or locate people and allow them to indicate they are safe[26]. Sometimes, data companies appear to possess capacities which states are deprived of. For instance, the French government has launched an application concurrent to Facebook's Safety check. Yet, it has proved itself as less efficient than Facebook's service[27]. Investments on data infrastructure and hiring the appropriate staff is thus necessary.

From a data architecture viewpoint, data economy is especially disruptive as most data companies process very large datasets that are not siloed. Hence, from seemingly trivial or technical data, services can emerge and potentially disrupt legacy organisations. On the contrary, public data are still often stored in siloed databases and public services still need to be digitalized so that citizens can easily interact with them.

Challenges

We identify three challenges when it comes to building a powerful European infrastructure for big data:

- **Data formats and standards** The exchange and processing of data relies on the ability to access and understand datasets. Common standards, on data formats, encryption, etc. are necessary to allow different actors to communicate and work together.
- **Public investment on data infrastructures** Data storage and processing facilities are at the heart of private data companies' - and of some states' - power.
- **Help to the transition to digital business models** Some sectors, such as taxis are disrupted by new players. Governments should help them transition to a data-based business model and protect their assets. Governments should also transition themselves to a digital model by becoming governments as a platform (O'Reilly 2011). Government as a platform refers to a radical architectural change in the infrastructure of governments which turn themselves into platforms on top of which citizens, or companies, may build services. It also refers to a change in the interaction between governments and their citizens and to the opening of data silos. There are examples of the building of such platforms in Lithuania[28], the United Kingdom[29] or

[25] http://www.cnbc.com/2014/12/11/la-residents-complain-about-waze-craze.html
[26] https://www.facebook.com/about/safetycheck/
[27] https://www.theguardian.com/world/2016/jul/16/nice-terroist-attack-france-saip-emergency-smartphone-app-failed
[28] https://lrv.lt/en/
[29] https://data.gov.uk/

France[30]. Turning governments into platforms require hiring the relevant staff or partnering with other actors. Governments should also design meaningful partnerships between data companies and governments. As we said, data companies are essential partners when it comes to governing. Instead of conflicting with them, we propose that governments try to develop meaningful partnerships which would prevent delegating complete prerogatives - such as the administration of dereferencing web links - data companies. It could also help governments to face the data transition and offer citizens new services. Eventually, such partnerships are necessary when it comes to accessing data for public interest.

Recommendations

Table 5. Action plan for data infrastructure.

	Short term actions	Actions to accomplish by 2030
Data formats and standards	Promote existing data standards	Europe leading in data standards definition
Public investment on data infrastructures	Define a strategy to improve public data capacities	European capacities rival with other areas such as the US or China
	Invest massively on innovative sectors	European research is a leader in sectors such as artificial intelligence.
	Invest massively on big data infrastructure and private sector	European companies are leaders in the big data sector
Help to the transition to digital business models	Survey the possibilities to digitalize disrupted sectors	Challenged sectors - such as taxies - are fully aligned with a digital business model.
	Assess the current state of digitalisation of governments	A European governmental platform is available

Table 5 provides an overview of the recommendations we see as necessary to build a big data infrastructure in Europe. The big data infrastructure implements the tools necessary for data governance and foster social good thanks to data economy.

This implementation relies on the definition possessing data and processing capabilities and making the necessary investments to do so ("Fact Sheet Data cPPP" 2014). It also relies on building a network of European data actors, and helping the development of existing ones while helping disrupted sectors to evolve in order to benefit from the data transition. This could help ease social tensions around the rise of the data economy.

[30] https://beta.gouv.fr/#cycle

3 BYTE RESEARCH ROADMAP

3.1 SCOPE AND METHODOLOGY

3.1.1 Roadmap purpose

One of the primary goals of the BYTE project is to devise a research and policy roadmap that provides incremental steps necessary to achieve the BYTE vision and guidelines to assist industry and scientists to address externalities in order to improve innovation and competitiveness. The roadmap, together with the community being built around it, focuses on giving good practice messages about societal issues in big data, and in particular to the environment, healthcare and smart city sectors, which have been selected by the BYTE big data community as the ones to be addressed first. The research roadmap focuses on the steps necessary for realising the most optimum use of big data (Cuquet and Fensel, 2016).

More specifically, the research roadmap focuses on what research, knowledge, technologies or skills are necessary in order to capture the economic and social benefits associated with the use of big data. It considers the positive externalities, negative externalities and social impacts associated with big data (Lammerant, De Hert and Lasierra Beamonte, et al. 2015) (Lammerant, De Hert and Vega Gorgojo, et al. 2015), maps research and innovation topics in the areas of data management, processing, analytics, protection, visualisation, as well as non-technical topics, to the externalities they can tackle, and provides a timeframe to address these topics and prioritisation of them. The research roadmap revolves around knowledge surrounding economic, legal, social, ethical and political issues, as well as standards, interoperability, development of meta-data, etc. It examines capacities and skills needed in computer science, statistics, social science and other industries or disciplines to enable European actors to take full advantage of the opportunities surrounding big data.

The BYTE roadmap is a key element of the process of building the BYTE big data community (BBDC), who will ultimately implement it to achieve the BYTE vision, continue the work of the BYTE project, investigating the positive and negative societal externalities of big data and the ways to make the best of them, and update the roadmap document yearly (Bigagli, et al. 2016). Such community especially targets civil society organisations, NGOs, the third sector, local governments, tech-transfer organisations and other non-profit organisations, as they are currently underrepresented in most big data fora and have so far been largely excluded from the EU Big Data debate (and resources). The BBDC will be autonomous, yet in strong synergy and collaboration with the Big Data Value Association (BDVA), a mainly industry-driven organisation officially endorsed by the European Commission as its private counterpart in the big data public-private partnership. Because of this collaboration with the BDVA, the present roadmap has been developed to be in alignment with the Big Data Value Strategic Research and Innovation Agenda (BDV SRIA) that defines the overall goals and technical and non-technical priorities for the European Public Private Partnership on Big Data Value (Big Data Value Association 2016). In particular, it addresses how the research and innovation topics of the BDV SRIA, amended and extended with the BYTE investigations and workshops contributions, can impact society (that is, how they can contribute to amplify positive externalities and diminish the negative ones). We also suggest incorporating the results of the present roadmap into the BDV SRIA, in particular to expand the societal part and non-technical priorities, which are currently unbalanced with respect to the technical ones.

3.1.2 Roadmap scope, and its relation to international Big Data roadmaps

The aim of this deliverable is to provide a research roadmap that presents what research and innovation is needed to capture positive externalities associated with big data and diminish

negative ones to obtain the best societal impact, develop the necessary skills and contribute to technology and data standardisation. It considers ***research and innovation in the five technical areas (data management, data processing, data analytics, data protection and data visualisation)*** used by the Big Data Value Association (Big Data Value Association 2016) and presents which topics have the highest priority to impact the societal externalities identified by the BYTE project in the upcoming 5 years, mid term (2025) and long term (2030). The roadmap is expected to guide European policy and research efforts to develop a socially responsible big data economy. We also expect to contribute to the Big Data Value Association activities and priorities by bringing a societal analysis of big data impacts, and to contribute to the creation of a multidisciplinary big data community around the BYTE results that includes as well NGOs, non-profit organisations, government (and especially local government) organisations, civil society organisations and citizens. Finally, and as part of the community building and dissemination efforts of the BYTE project, the roadmap and its conclusions will be fed into related European projects, such as the ones outlined in the appendix of (Bigagli, et al. 2016).

In this research roadmap, we have adopted a multi-layered approach (Phaal, Farrukh and Probert 2004) that accounts for skills development, standardisation and social impact of positive and negative externalities associated with big data, and links research and innovation topics to the targeted externalities and the sectors affected.

Being a cross-sectorial roadmap, another important aim of the present roadmap is to lead to action and collaboration among the BBDC members, who should adopt and update the roadmap after completion of the BYTE project. A strategy to achieve this has already been outlined in an interim report (Bigagli, et al. 2016), and has been further developed to produce a final version of the strategy for the BBDC. As we want to complement the Big Data Value Research and Innovation Agenda (Big Data Value Association 2016) and map how research will deliver the desired societal impact, we have adopted the same classification of research and innovation topics, which has been extended and further refined according to BYTE results and the stakeholders and community feedback. Another related side aim is to reach consensus within the community of the needs and the requirements to satisfy those needs. To this end, it is planned that the community will take up the task of monitoring and updating the roadmap after completion of the BYTE project, and that each year it will focus on three different sectors, starting from the ones studied within this project, to provide a deeper analysis of the needs and best practices in them. The first three sectors selected by the community have been the environment, healthcare and smart city sectors. An initial exploration of them is provided at the end of this roadmap, and further discussion will take place in the upcoming community workshop.

In the remaining part of this subsection we define ***the format that has been adopted for the research roadmap***. Phaal, Farrukh and Probert (2004) proposed a T-Plan fast-start approach to technology roadmapping that is primarily developed for use from a company perspective, but can be customised for a multi-organisational use of a group of stakeholders, and it has been explicitly done so in the context of disruptive technological trends. We followed it and complemented it with further guides especially tailored for sectorial technology roadmaps (see, e.g. (International Energy Agency 2014)).

The original time horizon for the present roadmap is 2020, to align with the Horizon 2020 objectives, although we have extended it to account for the upcoming 5 years after the roadmap presentation. In this roadmap, we have favoured a detailed timeframe of short term research priorities (2017-2021), with considerations as well in the mid (2025) and long term (2030).

We defined four top layers, sometimes labelled as *know-why* (Phaal, Farrukh and Probert 2004), that encapsulate the organisational **purpose** and correspond to the BYTE externalities

that the roadmap is intended to impact and potentiate (if positive) or diminish (if negative). The externalities are arranged in four areas and 18 coarse-grained externalities. In addition to these purpose layers, we also considered how this research impacts the different industry sectors studied in the BYTE project and beyond. These layers represent the society pull in the roadmap. We further defined six bottom layers, also known as *know-how*, corresponding to the five-technical and the non-technical research and innovation areas, or **resources**, that are to be addressed to meet the demands of the top layers, and that encode the technology push. Finally, the middle layers of the roadmap connect the purpose with the resources to **deliver** benefits to stakeholders, i.e. represents the *know-what*. This includes the skills development, standardisation efforts and societal impact that the research and innovation actions contribute to. Figure 10 sketches the architecture of the research roadmap.

	Now	2017	2018	2019	2020	2021	mid term	long term	time (know-when)
Economic externalities									purpose (know-why)
Social and ethical externalities									
Legal externalities									
Political externalities									
Industry sectors									
Societal impact									delivery (know-what)
Skills development									
Standardisation									
Data management									resources (know-how)
Data processing									
Data analytics									
Data protection									
Data visualisation									
Non-technical priorities									

Figure 10. Sketch of the research roadmap architecture.

The roadmap aims to deliver the vision for Europe. Nevertheless, to create it, we also took into consideration the Big Data research and innovation efforts and roadmaps from the whole world, and ***we are placing the roadmap in the international context***, emphasizing the key visible similarities and differences to the other countries of the world (Hajirahimova & Aliyeva, 2015). We relate to the three aspects of the roadmap: its construction process, the addressed research and innovation topics, as well as the prioritised public and private sector areas.

Roadmap construction process: We have constructed the roadmap applying similar instruments, like workshops, interviews, stakeholder consultations, as used elsewhere in the world. For example, details on the inputs for the US roadmap construction process are provided in its description (United States. Executive Office of the President, & Podesta, J., 2014). The literature that we used to construct the roadmap are internationally published publications, therefore all the possible topics relevant to Big Data are reflected. Similarly to the US and some

other roadmaps, national stakeholders have been consulted, in order to focus on the topics, which are most crucial for Europe, and where the European research and development is most competitive.

Research and innovation topics: Data Analytics direction is essentially present ubiquitously in the world's Big Data roadmaps. Especially, it tops the lists for the countries that have access to large amounts of data, such as for example the US, that have large quantities of the users' and companies' data from the Web, and Asian countries, that generally have access to massive amounts of data, originating from the Internet of Things. Machine learning and Deep Learning are also prominently present in a number of roadmaps, e.g. of USA, China (United States. Executive Office of the President, & Podesta, J., 2014; Huang, 2016): these research fields are related to data analytics and are relevant for applications such as recommendation and prediction. Open Data, its availability and role in making the public sector more transparent and efficient, have also substantially spread increased in the last years: it has been largely driven by the US and followed by developed countries on all continents. As large data brings large responsibility and eventually has an effect on individual lives of people, privacy and placing the users in control of their own data is mentioned throughout the roadmaps and acted on explicitly in the legislation base of the many of countries: USA (United States. Executive Office of the President, & Podesta, J., 2014), Russia[31], Japan[32], etc. Privacy-aware access to Big Data also has been identified as a high priority direction in our roadmap.

Industries: Many world's roadmaps, including ours, list specific industries/sectors, where the developments are most crucial and expected for the country for which the roadmap has been produced. While in our roadmap, the sectors of health, environment and smart city came in the first priority, other countries' priorities only are partly overlapping. Other countries have been identifying own industry, societal and public sector priorities. For example, the USA is explicitly supporting health care, education, homeland security, law enforcement and privacy law in public sector, as well as supporting the consumers and enterprises, advertising-supported industry, and data services in the private sector (United States. Executive Office of the President, & Podesta, J., 2014), while the most recent US Big Data case studies include access to credit, employment, higher education, and criminal justice (Munoz, 2016). And in China, for example, the manufacturing industry, as well as the environment and decrease of pollution, productivity of public sector and optimisation of transport are appearing in the high priorities (Huang, 2016; Hajirahimova & Aliyeva, 2015).

3.1.3 Roadmapping process

The roadmapping process built upon previous work within the BYTE project, and hence slightly deviated from common roadmapping method proposals and guidelines, which usually incorporate the creation of a vision and suggest conducting interviews with experts and holding several focus group discussions and workshops with relevant stakeholders. Such parts of the roadmapping process are already embedded in the BYTE project work and timeline and produced a number of deliverables that report the project findings and outcomes, the most

[31] "Proposed 'big data' law will empower Russians in the digital realm", Russia beyond the Headlines, 28 March 2017. https://www.rbth.com/news/2017/03/28/proposed-big-data-law-will-empower-russians-in-the-digital-realm_729263

[32] "The Internet economy and big data: Japan tackles data protection and privacy", Japan Today, 4 October 2015. https://japantoday.com/category/tech/the-internet-economy-and-big-data-japan-tackles-data-protection-and-privacy

relevant of which for the present roadmap are the horizontal analysis (Lammerant, De Hert and Lasierra Beamonte, et al. 2015), the evaluation of externalities (Lammerant, De Hert and Vega Gorgojo, et al. 2015) and the vision documents (Papachristos, Cunningham and Werker 2016) (Cunningham, et al. 2016). Thus, standard roadmapping processes were adapted to review such deliverables and incorporate their findings into the relevant parts of the roadmapping, rather than redoing the tasks.

The development process of the present research roadmap was done in three phases, adapting several recommendations and guidelines for the development of roadmaps to our case:[33]

1. Planning and preparation.
2. Visioning.
3. Development.

As roadmapping is a living process that incorporates the evolution of the roadmap after its finalisation, as well as its implementation, monitoring and updating (International Energy Agency 2014), a fourth phase is envisioned to be carried by the BYTE Big Data Community after completion of the roadmapping (and of the BYTE project as whole):

4. Implementation and adjustment.

In this regard, and as stated in a roadmapping guide, "the process is often as important as the resulting document, because it engages and aligns diverse stakeholders in a common course of action, sometimes for the first time" (International Energy Agency 2014, 4).

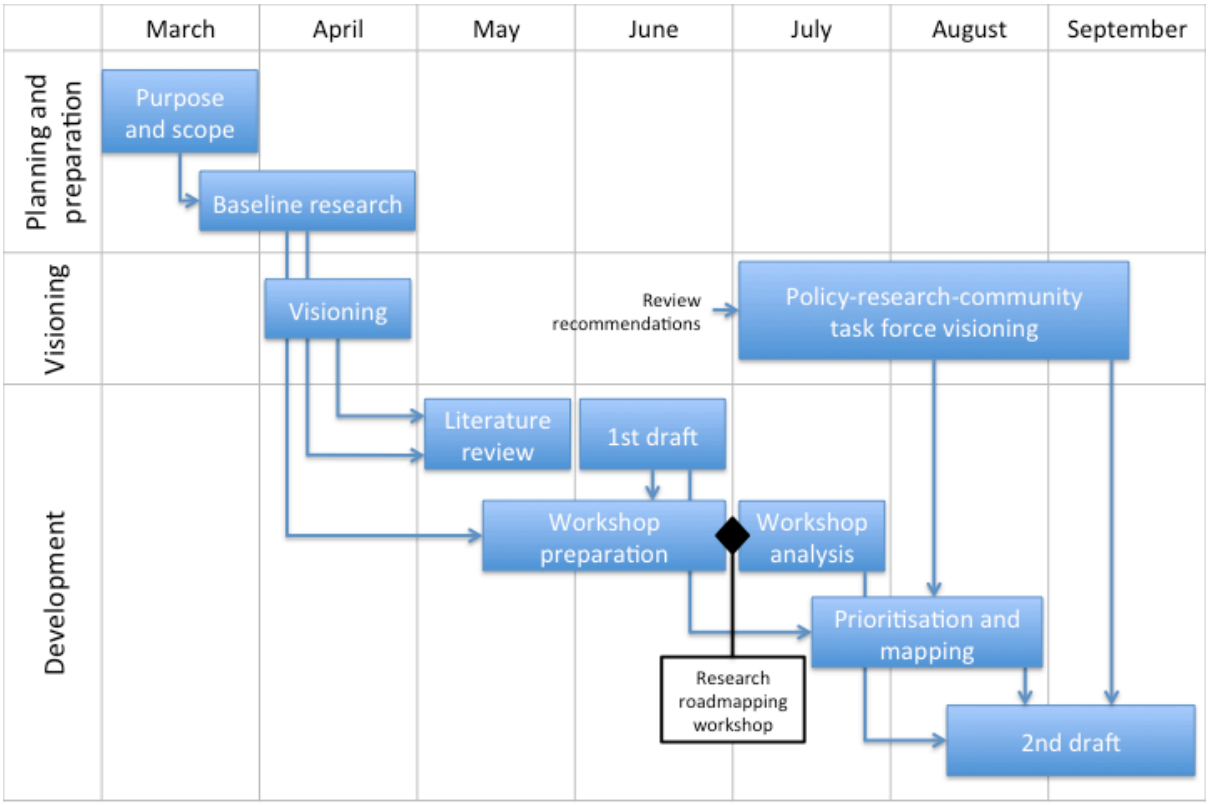


Figure 11. Research roadmapping phases and timeline. Not shown is the fourth and last phase (implementation and adjustment), expected to take place within the BYTE big data community formation and after the completion of the BYTE project.

[33] See for example (International Energy Agency 2014), (Phaal, Farrukh and Probert 2004).

Figure 11 depicts the timeline of these four phases and the associated tasks that were performed or are planned to take place in the future. The remaining of this subsection describes the four phases and their tasks.

Planning and preparation

In the first phase of the research roadmapping, a purpose and scope statement was developed to guide and maintain focus throughout the roadmap development process. Such purpose and scope was later refined to incorporate the harmonisation with the BYTE Big Data Community goals and development.

This phase included also a detailed plan of the roadmapping process, and addressed questions such as the boundaries of the roadmapping task, the areas (or layers) to be considered, the time frame for the roadmap, and the target audience.

Additionally, this phase identified stakeholders and relevant sectors beyond those studied by the BYTE project, in coordination with the BYTE vision recommendations and the community building process. We also reviewed the deliverables produced by the BYTE project so far to identify requirements and relevant research and innovation topics, with special emphasis on the ones affecting the research roadmap: D1.1, D2.1, D3.2, D4.1, D4.2, and D8.1.[34] Finally, a situation analysis was performed to determine if there was a lack of information in the BYTE deliverables. As it was recommended in the end of the BYTE vision document (Papachristos, Cunningham and Werker 2016, 55-57), the sectors studied in BYTE were complemented with extra ones. To this aim, it was especially useful to review the Big Data Technical Working Groups White Paper (Curry, et al. 2014) and its resulting roadmap (Becker, Jentzsch and Palmetshofer 2014) in the light of the BYTE externalities, the BDV SRIA research and innovation priorities (Big Data Value Association 2016), the Perspective on Big Data, Ethics and Society white paper (Metcalf, Keller and Boyd 2016), the NESSI white paper on big data (NESSI 2012), and the white paper Towards a Privacy Research Roadmap for the Computing Community (Cranor, et al. 2015).

The outcome of this phase is a general overview of the current big data research and innovation priorities in Europe, the societal externalities that can be tackled by them, and other resources (namely skills development and standardisation) that may be used to address them.

Visioning

A vision was already developed previous to the start of the roadmapping process and culminated in the *Foresight Workshop: Big Data Futures for Europe* and the definition of the BYTE vision, both presented in two deliverables that analyse future scenarios and recommend how to tackle the externalities of the vision (Papachristos, Cunningham and Werker 2016) (Cunningham, et al. 2016). In this second phase of the roadmapping process, the BYTE vision was summarised and clearly restated with a special focus in the topics of the research roadmap. This vision was subsequently amended to incorporate the project reviewers' recommendations and the discussions of the task force harmonising the goals of the policy and research roadmap and the community development.

[34] All deliverables can be found at the BYTE project website (http://byte-project.eu), as well as at the BYTE community in Zenodo (https://zenodo.org/communities/byte-eu/).

Development

In this phase, we used the externalities horizontal analysis (Lammerant, De Hert and Lasierra Beamonte, et al. 2015) and further developed the recommendations to address them (Lammerant, De Hert and Vega Gorgojo, et al. 2015). To this aim, we mapped how the research and innovation topics identified in the first phase may be used to address the societal externalities, and analysed how they may impact society and contribute to standardisation and skills development in order to capture the positive externalities. We also refined and extended the research challenges already outlined in BDV SRIA. This was done with a review of the documents mentioned in Section 0, plus additional ones as cited in the following sections.

To prioritise and bring the themes together on a time-basis (Phaal, Farrukh and Probert 2004), we conducted the ***BYTE Big data research roadmapping workshop*** on 1 July 2016, collocated with the European Data Forum 2016 that took place in Eindhoven, The Netherlands. The aim of the workshop was to present, discuss and obtain additional input from invited participants, in the format of round tables. It was a full day exercise with 26 participants from academia, industry, and non-governmental and non-profit organisations, as well as BYTE partners, advisory board members and BBDC founding members. In the workshop, the BYTE project and the research roadmap were presented, and three working sessions were held to discuss and validate the research topics and priorities, align them with their impact on externalities, and prioritise and place them in a time frame. These sessions were divided in 5 small round-table groups per session, each moderated by one BYTE partner. The attendees received all the relevant material prior to the workshop, including short descriptions of the project, the externalities, and the research topics and priorities being considered. The workshop also included an initial joint session with a parallel workshop organised by related EU projects Hobbit[35] and Big Data Europe[36] to help raise awareness of the BYTE and its research roadmap, the launch of the BYTE Big Data Community and a final wrap up together with the other two EU projects. The workshop programme and summary of participants can be found in Appendix 4.

Following this, we used the discussion and working groups of the roadmapping workshop to validate and further refine the priorities and research challenges, map them with their impact on societal externalities to complement the BYTE results and to temporally align and prioritise them. Finally, we incorporated and further expanded the action plan developed in the workshop.

The outcome of this phase is an itemisation of the research and innovation topics that need to be addressed and their expected impact on society, skills development and standardisation, their prioritisation and a timeline with an action plan to implement the roadmap.

Implementation and adjustment

The last phase takes place after finalisation of the roadmap, and will be taken up by the community building and the dissemination work packages. The roadmap will be launched via a press release, its distribution to the BYTE contact list, and public presentations by its leaders and other BYTE partners in relevant conferences and other events. The roadmap will be also presented in the community workshop, with a special focus on the environment, healthcare and smart city sectors, and at the final BYTE conference.

[35] http://project-hobbit.eu/

[36] http://www.big-data-europe.eu/

We intend this roadmap to be a living process, rather than stopping it after the publication of the present document. After the launch of the roadmap and its presentation to the BYTE community in the next community workshop, we envision that such community will take up the task to update it annually, each year with a special focus on three sectors. The community should therefore conduct expert workshops to monitor the progress of the roadmap implementation, reassess the research and innovation priorities and time alignment, and deliver best practices and recommendations especially tailored to the three sectors of the year. The community has already been engaged in roadmapping workshops and in the selection of the first three sectors.

3.2 REQUIREMENTS

3.2.1 Research and Innovation topics

In the development of the research roadmap, we have taken the approach outlined in the previous Sections 3.1.1 and 3.1.2, and have considered the research and innovation topics of the BDVA's Strategic Research and Innovation Agenda (SRIA). These topics have been further extended with observations and recommendations from the BYTE case studies, analysis and workshops and the contribution from community stakeholders, and aligned (in sections below) with the impact they have in society (i.e. how can they amplify the positive externalities and diminish the negative ones). The aim is that the roadmap should later be incorporated into the BDVA SRIA, in particular to expand the societal part and non-technical priorities, which are currently unbalanced with respect to the technical ones.

In Table 6 and Table 7 the research topics and non-technical priorities currently being considered at the BDVA are listed. These 5 technical and 1 non-technical priorities were selected following three phases: the identification of the most important challenges via a structured needs and requirements analysis in a series of workshops focused on specific sectors, the clustering and mapping of these challenges to the roles of the big data value chain, and the alignment with existing big data solutions (Big Data Value Association 2016, p. 21-23). There, the analysis of the priorities was focused on the technical needs to turn big data into value, the most important of which were: data integration and harmonisation across sources; data curation, veracity and their life-cycle; low latency and real-time processing; advanced analytics; data protection and privacy; and advanced visualisation, user experience and usability (Big Data Value Association 2016, p. 21).

A comprehensive background of the data management, data processing, data analytics, data protection, data visualisation and non-technical priorities is given in the BDV SRIA document (Big Data Value Association 2016). In this section, we complement this background and challenges with a focus on the requirements that have a societal impact by a further review of the literature complemented with the further feedback and validation from the research roadmapping workshop, as described in the methodology section. Thus, we add only observation and subtopics not present in (Big Data Value Association 2016). These requirements are then used in Section 3.2.3 below to prioritise and identify their impact in sectors and societal externalites.

Table 6. Research and innovation topics (1/2: Data management, data processing and data analytics).

Data management	Data processing	Data analytics
Handling unstructured and semi-structured data	Architectures for data-at-rest and data-in-motion	Improved models and simulations
Semantic interoperability	Techniques and tools for processing real-time heterogeneous data	Semantic analysis
Measuring and assuring data quality	Scalable algorithms and techniques for real-time analytics	Event and pattern discovery
Data lifecycle	Decentralised architectures	Multimedia (unstructured) data mining
Data provenance, control and IPR	Efficient mechanisms for storage and processing	Machine learning techniques, deep learning for BI, predictive and prescriptive analytics
Data-as-a-service model and paradigm		Context-aware analytics

Table 7. Research and innovation topics (2/2: Data protection, data visualisation, non-technical priorities).

Data protection	Data visualisation	Non-technical priorities
Complete data protection framework	End user visualisation and analytics	Establish and increase trust
Data minimization	Dynamic clustering of information	Privacy-by-design, security-by-design, anti-discrimination-by-design
Privacy-preserving mining algorithms	New visualisation for geospatial data	Ethical issues
Robust anonymisation algorithms	Interrelated data and semantics relationships	Develop new business models
Protection against reversibility	Qualitative analysis at a high semantic level	Citizen research
Pattern hiding mechanism	Real-time and collaborative 3-D visualisation	Discrimination discovery and prevention
Secure multiparty mining mechanism	Time dimension of big data	
	Real-time adaptable and interactive visualisation	

Data management

Handling unstructured and semi-structured data. It was pointed out by stakeholders at the workshop that multilingualism is still a challenge, especially in the processing of natural languages different from English. It has also been identified as an opportunity for Europe, as it already has the necessary skillset to address these requirements. There is also a need to develop easy-to-use reporting tools including semantic annotations that do not add extra work, e.g. to healthcare professionals (Lyko et al. 2016). This requirement is relevant in conjunction with the topic below.

Semantic interoperability. A relevant subtopic to be added is data integration and fusion, as pointed out by stakeholders. There are issues with format conversion that lead to intelligence

losses. Currently, reengineering is the only way to recover intelligence, so there it is thus required to define new policies about how original files are kept, as well as to develop technology to ensure interoperability among different formats. In any case, linked data technologies need to be simplified in order to make them easily adoptable (Domingue et al. 2016). Also, semantic search, schema matching and mapping, and ontology alignment have to be addressed (Freitas and Curry 2016).

Measuring and assuring data quality. This topic should include transparency on the data collection process, and also meta information on the context and purpose of such collection. Approaches to compute the uncertainty of the results of algorithms need to be developed (Freitas and Curry 2016), in order to include evidence-based measurement models of uncertainty over data. Also, algorithms to validate and annotate data need to be developed (Freitas and Curry 2016). Finally, funding agencies should require an explicit estimate of data curation and publication costs of high quality data (Freitas and Curry 2016).

Data lifecycle. Access to data is still put forward as one of the main challenges. It was mentioned that focus should be given on data that already exists rather than on data that needs to be created in the future, and thus data creation, with a focus on surfacing already existing data, is a priority within this topic. There is more data out there than what people realise, and it should be made easier to find (Domingue et al. 2016). This could be assisted by developing search engines for datasets with ranking, in order to drive owners to publish better datasets, following the improvement of websites that want to appear high in Google rankings (Domingue et al. 2016). Adaptive data detection and acquisition is needed in e.g. the finance and insurance sector (Lyko et al. 2016). Other relevant subtopics are data discovery, datasets crawlers, metadata, dataset ranking (Domingue et al. 2016). Data curation by demonstration, in analogy to programming by example or by demonstration, would also for the distribution and scalability of the system. Also to be added here is the preservation and archiving of data. It has also arisen from several stakeholders that currently the biggest challenge is the variety of data. Within this topic, data citation, curation and preservation have been identified as additional relevant subtopics. In particular, standards for data citation are currently demanded. Understanding how data expires, what happens with historical data and how it is archived is important. In relation to this, the synchronisation of data and how to update extracted knowledge bases if the sources are changing should also be addressed (Lyko et al. 2016). Open data is also central to research itself. The Research Data Association is actively addressing how to build research data services with open linked data, for example in the publishing, referencing, citing and searching areas (Peroni et al. 2015, Thanos 2016). There are promising semantic languages and technologies that can be applied to such research data services (e.g. linked services, linked data, Schema.org), which can as well assist in multichannel research dissemination (Gruber 1993, Guarino, Poli and Gruber 1993, Fensel 2014).

Data provenance, control and IPR. It has been highlighted that certain kinds of data attract new rights and require new rights statement initiatives. This has also a further impact in its implications within linked open data, and is particularly relevant for media data, where the digitalisation of an object (other than text) was put forward as an example. Data licensing and ownership have still no means to be represented clearly for everybody, and this is especially dangerous in Internet of things applications, where data is distributed among different physical locations and where often the appliance and software manufacturers are the organisations that

grab the data[37]. Data curation depends on mechanisms to assign permissions and digital rights at the data level and to provide context through data provenance (Freitas and Curry 2016). New theoretical models and methodologies for data transportability under different contexts should be developed (Freitas and Curry 2016). Decisions taken during the data curation process need to be captured, and models and tools to grant fine-grained permission management developed (Freitas and Curry 2016). In this same direction one finds sandboxing and virtualisation techniques (Strohbach et al. 2016). Nanopublications may impact the development of distributed science via semi-structured data and scientific statements (Groth, Gibson and Velterop 2010). Another open challenge is how to guarantee the integrity and confidentiality of provenance data (Strohbach et al. 2016). There are also many policy issues related with this topic, which are addressed in the policy part of the roadmap.

Data-as-a-service model and paradigm. Relevant subtopics are licensing, ownership and marketplace. Research has to be devoted also to the extraction of value from data, particularly in terms of what data needs to be created for maximum value extraction. In this regard, also the estimation of data value both in the present and future is a relevant topic. More services that take advantage of open data need to be supported (Lammerant, De Hert, Vega Gorgojo et al. 2015, 12). New distributed techniques, such as blockchain, are in the process to be established and adopted for different sectors, including education (English et al., 2016).

Also within the area of data management, and in relation to **open data practices**, it has been pointed out that there exist several open data issues for registered companies and lack of harmonization across Europe. This is not restricted only to the legal and policy framework, but can be addressed also from a technical perspective. In general, industry agrees with the need to open data but finds difficulty to make it open (especially in orphan works), and asks for financial support to open data. We thus recommend adding an **open data and data creation** priority within the area of data management aimed at surfacing already existing data. Public funding should keep prioritising processes that support open data initiatives (Domingue et al. 2016). Tracking and recognition of data and infrastructure should be improved in academic research (Freitas and Curry 2016). Options for this include recognising open dataset publication analogously to paper publication in journals. Good practices would also include actions for assisting the publication of high quality open data to organisations with fewer resources (manpower, skills, money, etc.), as it is often essential to their competitiveness –see for example, the issues with real life open data available in the tourism sector (Kärle et al. 2016)

Data processing

Techniques and tools for processing real-time heterogeneous data. This is particularly needed in the development of new tools for sensor data processing, especially in the manufacturing, retail and transport sectors (Lyko et al. 2016), as well as in the energy sector (Simsek et al. 2016). Such technology is often encountered in the area of Internet of Things, and domains like smart cities. Here, due to the dense and numerous population (i.e. more data) and pressing needs, the developed countries in Asia may eventually become the world leaders in most cases. In particular, in the telecommunications sector, Asia possesses the whole institutions partly working on topics that are out of scope for the EU mobile operators, due to their limited subscriber bases –for example, innovative end user mobile services (Qiao et al. 2015). Social media mining is also relevant in this area (Lyko et al. 2016).

[37] Semantic data licensing approaches are though currently in development e.g. at W3C, and in ongoing projects such as DALICC: https://www.dalicc.net

Scalable algorithms and techniques for real-time analytics such as stream data mining in contexts of a high volume of stream data coming from e.g. sensor networks or large numbers of online users (Domingue et al. 2016), as well as real-time analysis of public transportation data (Lammerant, De Hert, Vega Gorgojo et al. 2015, 10).

Decentralised and distributed architectures. This includes efficient and scalable cryptographic mechanisms for the cloud (e.g. directory-based encryption, container-based encryption, manual encryption) and attribute-based encryption (Kamara and Lauter 2010, Strohbach et al. 2016) . Distributed architectures should also be explored as an opportunity to keep sensitive data on user-governed devices.

Efficient mechanisms for storage and processing. To automate complex tasks and make them scalable, hybrid human-algorithmic data curation approaches have to be further developed (Freitas and Curry 2016). This research and development sub-area has been developing rapidly in the last years, delivering new types of massive data storage and processing products e.g. NoSQL knowledge bases. Basing on the advances of cloud computing, the technology market is very developed in this area, with most knowledge bases products created in the US (see Sharma, 2016, for an overview). Crowdsourcing also plays an important role. Energy-efficient data storage methods are also a crucial research priority (Strohbach et al. 2016).

Data analytics

Improved models and simulations. There is a need of better integration between algorithmic and human computation approaches (Freitas and Curry 2016). Catchment techniques, recommendations and customer tendency research are also relevant for the retail sector (Domingue et al. 2016, Lammerant, De Hert, Vega Gorgojo et al. 2015, 11), and simulations for resource allocation for the crisis informatics and smart city sectors (Lammerant, De Hert, Vega Gorgojo et al. 2015, 10). In general, most models would extremely benefit by methods to correct sample bias (Becker 2016). This affects also the data collection and quality assessment processes.

Semantic analysis. Examples are sentiment analysis, a relevant subtopic when using social media data for the manufacturing or retail sectors (Lyko et al. 2016), and entity recognition and linking (Freitas and Curry 2016).

Event and pattern discovery. To be added within this innovation topic, but also relate to the predictive and prescriptive analytics below, is the need to further investigate and differentiate between correlation and causation. In this direction, an evidence-driven, bottom-up approach has been put forward by (Brodie 2015) to first deduce correlations from evidence (eg using data from economic phenomena) and then develop means to estimate their correctness and completeness, such as the probabilistic likelihood that correlations are causal within error bounds. Anomaly detection can be applied e.g. to detect deviations from traffic in a smart city (Domingue et al. 2016). Clustering of social media post can also be used to detect and gather real-time information in emergencies (Domingue et al. 2016), especially in real time (Becker 2016). Another topic is pattern recognition on imaging device results (Lammerant, De Hert, Vega Gorgojo et al. 2015, 10).

Multimedia (unstructured) data mining. Sentiment analysis beyond the analysis of textual information needs to be addressed (Lammerant, De Hert, Vega Gorgojo et al. 2015, 10). In this regard, it was raised during the workshop that there is a lack of tools to deal with multilingual sentiment analysis, and Europe is probably in the best position to tackle this challenge.

Machine learning techniques, especially deep learning for business intelligence, predictive and prescriptive analytics. This topic is already coupled in the BDV SRIA document with the priorities on visualisation and end-user usability. But even more importantly that such usability of the analytics results by non-data scientists, it has been recognised an urgent need of validated methodologies and standards behind the analytics on whose results decisions are to be taken, and that are easily identifiable and understandable by decision-makers. Also relevant for decision-making is to correctly assess the representativeness of data and possible data biases, as it may lead to biased decisions. An emerging trend is to use new sensor data for predictive analysis, e.g. in Industry 4.0 (Becker 2016). This field has been particularly strong in the US, showcased by such recent developments as IBM's Watson or Google's AlphaGo (Silver et al., 2016).

Data protection

Complete data protection framework. It is important to create "resources for using commoditized and privacy preserving Big Data analytical services within SME's" (NESSI 2012, 25). The major security challenges are now in non-relational data stores (Strohbach et al. 2016). Also granular access controls have to be developed that allow sharing data on a fine-grained level (Freitas and Curry 2016). New legal means have to be developed too to handle access to data and the permitted use of data that ensure that data protection is not an obstacle for big data practices (Lammerant, De Hert, Vega Gorgojo et al. 2015, 72). This includes a better scalable transaction model in data protection law (Lammerant, De Hert, Vega Gorgojo et al. 2015, 90).

Privacy-preserving mining algorithms. Although further research is still needed in this area, there exist already interesting approaches that are however not well known in industry. There is a need thus to disseminate these results and bring them to practice (NESSI 2012, 25). Special emphasis has to be done in the mining algorithms of social media (Lammerant, De Hert, Vega Gorgojo et al. 2015, 10).

Robust anonymisation algorithms. This includes the development of novel algorithms such as k-anonymity (Sweeney 2002).

Protection against reversibility. Considerable research is required to better understand how data can be misused, how it needs to be protected and integrated in big data storage solutions (Strohbach et al. 2016).

Data visualisation

End user centric visualisation and analytics. Natural language interfaces, and interactive and easy-to-use data access and transformation methods need to be further developed and brought to commercial applications (Freitas and Curry 2016).

Dynamic clustering of information. This requires efforts for new and better data summarisation and visualisation, and user interfaces for parallel exploration (Becker 2016) such as subjunctive interfaces (Lunzer and Hornbaek 2008).

New visualisation for geospatial data. Also geospatial data can benefit from the user interfaces for parallel exploration mentioned above (Becker 2016).

Interrelated data and semantics relationships. Semantic search.

Qualitative analysis at a high semantic level. Intuitive data transformation interfaces.

Non-technical priorities

Establish and increase trust. Open government data is widely recognised as a method to increase trust and transparency (Domingue et al. 2016), although this requires parallel actions to reduce the digital divide (Lammerant, De Hert, Vega Gorgojo et al. 2015, 30-31). A solution is an increase of data journalists who are able to process and present such data to a wider audience.

Privacy-by-design. Transparency for users is still an issue, so privacy-by-design and similar by-design approaches are vital (Domingue et al. 2016). By-design approaches are generally seen as a solution to allow business to evaluate and analyse data, and in particular sensitive data, without needing too restrictive provisions to avoid profiling (NESSI 2012, 25). An example is the fine-grained control of digital rights (Qin and Atluri 2003). As anonymising and de-identifying data might be usually insufficient in view of the amount of data that can be used for re-identification, the transparent handling of data and algorithms and company audits should be considered (Strobach et al. 2016).

Ethical issues. It has also been pointed out that further discussion is needed regarding whether research that analyses human data should fall within the regulations of research based on human subjects. This is in line with the discussions presented by the Council of Big Data, Ethics and Society (Metcalf, Keller and Boyd 2016, 16).

There is a demand from the scientific community to access data owned by companies for research purposes. Standards should be set to enable such sharing of data across sectors in a way that allows companies to contribute anonymised data to the scientific community without the possibility of backfiring, as has happened in past experiences (Metcalf, Keller and Boyd 2016, 15-16).

Research is needed to quantify the risks posed by data science practices that rely on big data. This includes dealing with minimal individual risks that however affect a very large population and with privacy risks that depend on highly varying privacy expectations of subjects in the same study (Metcalf, Keller and Boyd 2016, 17).

Research is also needed to account for and mitigate the risks of sharing datasets that can be later combined with auxiliary datasets, thus e.g. increasing the risk of de-anonymisation. Research has already started in this direction (Wan et al. 2015).

Usage of publicly available, although illicitly obtained data sets is also a matter of controversy within the scientific community (Metcalf, Keller and Boyd 2016, 18). There is a need to establish at least best practices on how to approach this challenge.

In industry, ethic processes and ethic review structures that work have to be developed and tested (Metcalf, Keller and Boyd 2016, 18).

Bias is also a relevant ethical issue that needs further research. It is commonly implicit in big data processes that all data will eventually be sampled, although this is hardly ever true and there can indeed be a sample bias introduced by technical, economic or social factors (Becker 2016). Subjective bias can also be introduced in the data through the labelling of the data (Domingue et al. 2016).

Develop new business models. Open source big data analytics have been proposed as a way to ensure that benefits remain in the EU (NESSI 2012, 24). However, moving beyond the Open Data Initiative to an interoperable data scheme to process data from heterogeneous sources is also seen as a way to foster and develop new business models (NESSI 2012, 25). Other novel models are pre-competitive partnerships where organisations that are typically competitors

cooperate in R&D projects of certain data value chain steps, such as data curation, that do not affect their competitive advantage and public-private partnerships (Freitas and Curry 2016).

Citizen research. Crowdsourcing may be used to increase data accuracy (Domingue et al. 2016) and scale data curation (Freitas and Curry 2016), among other applications. New methods are needed to route tasks to crowdsourcing participants based on their expertise, demographic profiles, and long-term teams, and develop open platforms for voluntary work (Freitas and Curry 2016). Research is needed to better understand the social engagement mechanisms, e.g. in projects such as Wikipedia, GalaxyZoo (Forston et al 2012) or FoldIt (Khatib et al 2011), which would amplify community engagement (Freitas and Curry 2016).

Discrimination discovery and prevention. Within this topic, or as a priority of its own, more research on legal informatics and algorithm accountability is needed. This is especially relevant for IPR-related externalities.

3.2.2 Externalities

The BYTE project identified and considered 73 societal externalities classified by the pairs of stakeholders involved (public sector, private sector and citizens) and their main topic (business models, data sources and open data, policies and legal issues, social and ethical issues, and technologies and infrastructures). The full list can be found in the appendix of the Case study reports (Vega-Gorgojo, Donovan, et al. 2015, 152-154). Throughout the project, we have used the following definition for externality (Vega-Gorgojo, Løvoll, et al. 2014, 9):

- Positive externalities occur when a product, activity or decision by an actor causes positive effects or benefits realized by a third party resulting from a transaction in which they had no direct involvement.
- Negative externalities occur when a product, activity or decision by an actor causes costs (or harm) that is not entirely born by that actor but that affects a third party, e.g. society. It is generally viewed as a failure of the market because the level of consumption or production of the product is higher than what the society requires.

As the boundary between internal and external is often arbitrary, we have in some cases extended the definition to include also the internal impact of a product, activity or decision.

Table 8. Overview of societal externalities by their main area.

Economic	Social and ethical	Legal	Political
Improved efficiency	Improved efficiency and innovation	Privacy	Private vs. public and non-profit sector
Innovation	Improved awareness and decision-making	IPR	Losing control to actor abroad
Changing business models	Participation	Liability and accountability	Improved decision-making and participation
Employment	Equality		Political abuse and surveillance
Dependency on public funding	Discrimination		
	Trust		

The 73 externalities were simplified to 18, and grouped in four main areas: 1) economic externalities, 2) social and ethical externalities, 3) legal externalities, and 4) political

externalities (Lammerant, De Hert and Lasierra Beamonte, et al. 2015, 28-30), as shown in Table 8.

In what follows, we present a short description and list of the externalities associated to each group. We do not extend further, as they have already been thoroughly discussed in previous BYTE reports, including how they show up in different sectors (Donovan, et al. 2014, Vega-Gorgojo, Donovan, et al. 2015, Lammerant, De Hert and Lasierra Beamonte, et al. 2015).

Economic externalities

Improved efficiency of existing processes that resulted from big data activities. Better and more targeted services through data sharing, analysis and profiling. Cost-effectiveness of services. Gather public insight by identifying social trends and statistics. Lack of context or incomplete data. Increased demand in computing power, data storage or network capabilities.

Innovation, that is formation of new value chains or major transformation of existing ones. Better services through data sharing and analysis. Foster innovation from government and open data. Accelerate scientific progress and data-driven R&D. Economic growth and innovative business models through community building and open data. Reduced innovation due to restrictive legislation.

Changing business models, for example in the development of new services based on a new use of data or by specialisation in specific services needed as part of the data value chain. Economic growth and innovative business models through community building and open data. Challenge of traditional non-digital services. Competitive disadvantage of newer businesses and SMEs. Creation of a few dominant market players. Inequalities to data access. Open data puts the private sector at a competitive advantage. Employment losses for certain job categories. Privatization of essential utilities.

Employment. Data-driven employment offerings. Employment losses for certain job categories. Opportunities for economic growth through open data.

Dependency on public funding, especially to kick-start a data economy. Open data puts the private sector at a competitive advantage. Innovative business models through community building.

Social and ethical externalities

Improved efficiency and innovation that produce social benefits. Tracking environmental challenges. Better and more targeted services through data sharing, analysis and profiling. Crime prevention and detection. Safe and environment-friendly operations.

Improved awareness and decision-making that produce social benefits. Gather public insight by identifying social trends and statistics. Tracking environmental challenges. Better and more targeted services through data sharing, analysis and profiling. Safe and environment-friendly operations. Crime prevention and detection. Transparency and accountability of the public sector.

Participation triggered by big data practices. Increased citizen participation. Support communities.

Equality to benefit from big data, for example due to lack of skills, access, infrastructure or language. Discriminatory practices and targeted advertising. Distrust on data coming from uncontrolled sources.

Discrimination based on the use of data for profiling, concerns about political abuse and prosecuting specific groups, etc. Increase awareness about privacy violations and ethical

issues of big data. Discriminatory practices and targeted advertising. Private data misuse, especially sharing with third parties without consent.

Trust problems that create a barrier due to e.g. concerns about exploitation and manipulation, privacy violation or data abuse. Increase awareness about privacy violations and ethical issues of big data. Public reluctance to provide information. Consumer manipulation. Lack of context or incomplete data. Market manipulation. Distrust on data coming from uncontrolled sources. Private data misuse. Continuous and invisible surveillance.

Legal externalities

Privacy and data protection. Private data misuse. Threats to data protection and personal privacy. Private data accumulation and ownership. Increase awareness about privacy violations and ethical issues of big data. Invasive use of information. Discriminatory practices and targeted advertising. Consumer manipulation. Continuous and invisible surveillance. Reduced innovation due to restrictive legislation. Barriers to market entry.

IPR. Threats to intellectual property rights. Private data accumulation and ownership. Lack of norms for data storage and processing. Reduced innovation due to restrictive legislation. Need to reconcile different laws and agreements.

Liability and accountability. Lack of norms for data storage and processing.

Political externalities

Private vs. public and non-profit sector. Dependency on external data sources, platforms and services. Competitive disadvantage of newer businesses and SMEs. Privatization of essential utilities.

Losing control to actors abroad. Privatization of essential utilities. Political tensions due to surveillance out of the boundaries of states. Need to reconcile different laws and agreements. Lack of norms for data storage and processing. Increased transparency.

Improved decision-making and participation. Transparency and accountability of the public sector. Support communities. Increased citizen participation.

Political abuse and surveillance. Political tensions due to surveillance out of the boundaries of states. Private data misuse, especially sharing with third parties without consent.

Externalities consolidation

The relevance of each externality to the BYTE sectors have been assessed by a review of the case study reports and complemented with an analysis of big data initiatives and external studies. This relevance is summarised in a heat map in Figure 12. The included BYTE sectors are crisis informatics, culture, energy, environment, healthcare and smart city (see e.g. *Big data for good* (Cuquet, Vega-Gorgojo, et al. 2016) for an overview of the cases and their associated externalities and (Cuquet, Vega-Gorgojo, et al. 2017)). As recommended in the BYTE vision and foresight analysis (Papachristos, Cunningham and Werker 2016, 55-57), we have extended the analysis of big data societal impact to new sectors. To do so, we have given a special focus on those sectors previously analysed by the BIG project, mainly from a technical perspective (Curry, et al. 2014), to allow for an alignment of research and innovation topics with their societal relevance. These sectors are administration, education, finance and insurance, logistics, manufacturing, maritime, media, retail, science, telecommunication, tourism and transport. The results are displayed in Figure 21 (Appendix 3).

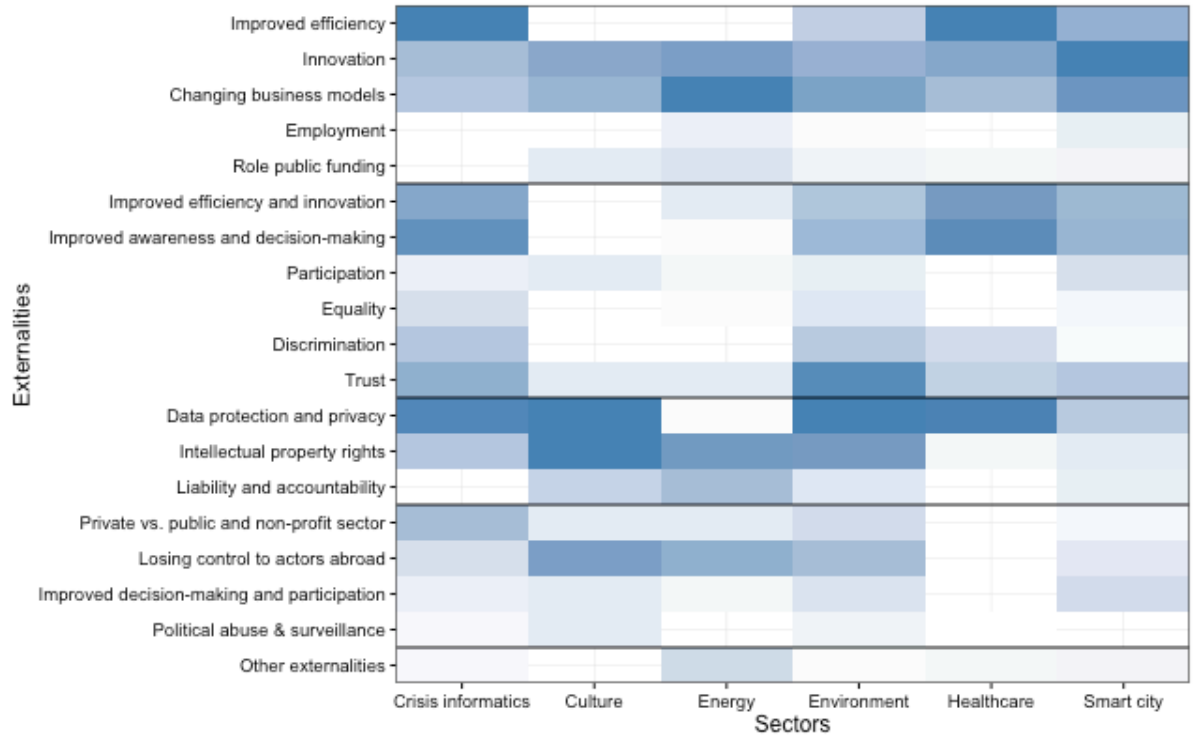


Figure 12. Relevance of externalities in the BYTE sectors, derived from BYTE analysis and the literature review. Relevance has been normalised by sector: darkest blue corresponds to the most relevant externality in the sector.

Other non-technical priorities

In addition to the BYTE externalities, several non-technical priorities arising from the BDV SRIA, the BYTE results and the workshop discussions that can be address by research and innovation actions are considered here. These priorities can be grouped into skills development and standardisation efforts.

The need for educated people equipped with the right data skills has been extensively identified (see e.g. Manyika, et al. 2011, NESSI 2012, Mattmann 2013, e-skills uk 2013, Curry, et al. 2014, 32, Berger, et al. 2014, 50-51). For example, the McKinsey report classifies the required skills in *deep analytical talents* to analyse the data, *data-savvy managers and analysts* to effectively consume the data and *supporting technology personnel* (Manyika, et al. 2011). Similarly, the BDVA has also identified three profiles that partially overlap with the ones previously described: *data scientists*, *data-intensive business experts* and *data-intensive engineers* (Big Data Value Association 2016, 33). The European Data Science Academy project[38] is addressing this challenge and has recently released a report that evaluates the skills gap and how to close it (Mack, Tarrant and Dadzie 2016).

The BYTE case studies and analysis also identified and confirmed this need and recommend promoting big data in education policies (Lammerant, De Hert, Vega Gorgojo, et al. 2015, 11, 23-24). This has been further confirmed in the big data research roadmapping workshop. It was noted that an interface between policy makers, society and industry is needed. This requires data-savvy professionals in all areas that have to take data-driven decisions, and not only the ubiquitously stated need of data scientists. Moreover, stress has been put to integrate ethics education into the data science curricula. This is supported by similar recommendations elsewhere (Metcalf, Keller and Boyd 2016, 13). In the research area, priority should be given

[38] https://edsa-project.eu

to simplify already mature technologies (Domingue et al. 2016) and make them accessible to innovative businesses. In successive workshops, it has repeatedly been mentioned that the increase in data skills in the general public and in key expert positions could mitigate the large need of data scientists and engineers. Data-intensive policy makers are an example of a skill that was identified to be of high priority: more than the ability to deal with data, that of being able to correctly understand and interpret the models used to predict and make recommendation, and the type of data in which they are based. This includes more research into correlation vs. causation tools and the need of validated methodologies.

Another relevant aspect that came up is the digital divide and how open data, that supposedly benefits citizens in general, is actually more likely to increase the digital divide and produce social inequality, as data is effectively data "is only open to a small elite of technical specialists who know how to interpret and use it", and to those who can employ them (Lammerant, De Hert and Vega Gorgojo, et al. 2015, 30) (Roberts 2012). This divide also affects the gender category: female leaders in industry and research are only a small percentage (e.g., 11% in ICT in Austria, compared to an already low 25% of average in other sectors, Berger et al. 2014, 51). An important action to decrease this divide is to promote data journalism to process, digest, and present the newly available open data to society.

Regarding standardisation, in general two types have been identified in alignment with the BDV SRIA: *technology* and *data standardisation*. It was though pointed out that an excess of standards, especially for interoperability, is not always useful and can lead to potentially negative changes in society, especially when they slow down innovation. In this case they should be replaced by best practice recommendations. Standardisation is in any case urgently needed for *data citation*. Other identified requirements have been vocabulary standardisation and the need of open APIs. For example, data and conceptual model standards (e.g. ontologies and vocabularies) strongly reduce the data curation effort and simplify data reuse (Freitas and Curry 2016). The development of minimum information models following the example of MIRIAM (Le Novère and Laibe 2007) would improve data curation. Query interfaces are also in need of a standard (Strohbach et al 2016).

3.2.3 Prioritisation and mapping

The research and innovation topics of Section 3.2.1 have been mapped to the societal impact they can have in terms of the economic, social and ethical, legal, and political externalities presented in Section 3.2.2 and to the sectors considered in BYTE and in the externalities consolidation within Section 0. The mapping has been done via a review of the BYTE studies and external resources investigating technical requirements, mainly (Cavanillas, Curry, and Wahlster 2016) as this is the main resource from where the BDV SRIA document has evolved. In parallel, the mapping has been done as well at the research roadmapping workshop, where the research topics were also prioritised, extended and revised. Figure 13 summarises the relevance of research and innovation topics for each of the sectors. Figure 14 evaluates the impact of these topics in the societal externalities, and finally Figure 15 shows an independent evaluation of such mapping by the industry and academia experts that participated in the research roadmapping workshop. In the workshop, participants were arranged in groups and codified the priority as high, medium, low or none. Here, we have aggregated the contributions and used a colour scheme from dark blue to white to code the priority of each topic to:

Top priority (darkest blue) if all or almost all stakeholders agreed the topic to be of high priority.

High priority if it was generally considered to be of high priority.

Medium priority (if it was generally considered to be of medium priority.

Low priority otherwise.

No priority (in white) if all stakeholders agreed the topic has no priority

As it can be seen by comparing Figure 14 and Figure 15, the findings of the literature review and the workshop are in good agreement.

To increase the likelihood of technology adoption in the future, the following considerations regarding research topics and their social impact were also put forward by the community in the research roadmapping workshop.

The area of data management is of high priority:

- Without management, there is **restricted efficiency** and **low economic output**.
- The **data lifecycle** is co-dependent on **public funding**.
- New business models can be created within the **data-as-a-service paradigm**, such as paying for data cleaning.
- **Discrimination** and **trust** are strongly affected by **data management topics**, especially when participation is diminished. Citizens trust is increased by better **data provenance, control and IPR**. On the other hand, businesses trust via innovations in **data-as-a-service model and paradigm**.
- There exists also the risk of **losing data to models abroad** caused by an **inefficient data management**.
- **Structured and semi-structured data** are inefficient, especially when processed.

The area of data processing has moderate to high priority:

- For data-based policy making, it should be considered enforcing at least 3 stars (make data available in a non-proprietary open format)[39]. However, it's also worth mentioning that continuing with the current 1-star assessment opens opportunities for other players to create 3-star processing products.
- **Real time efficiency** has moderate to high priority, but legal and especially trust issues require clarification.
- The debate between **decentralized and centralised architectures** needs to be decided. For example, **participation** may be positively affected if decentralised, while it was mentioned that the only way for purely **open data** is centralization.

The area of data analytics has also moderate to high priority:

- Research into **multimedia data mining** will lead to **new business models** and innovation, but has important **IPR issues**. Most licenses do not allow for data mining, but the development of **blockchain** may lead to higher participation thanks to an increase of **trust**.
- Smart contracts are a priority.
- Within machine learning techniques for business intelligence, advances in **auditing algorithms** will have a positive impact in **equality**, **discrimination** and **trust** externalities, as well as in **liability and accountability**.

[39] See http://5stardata.info for more details.

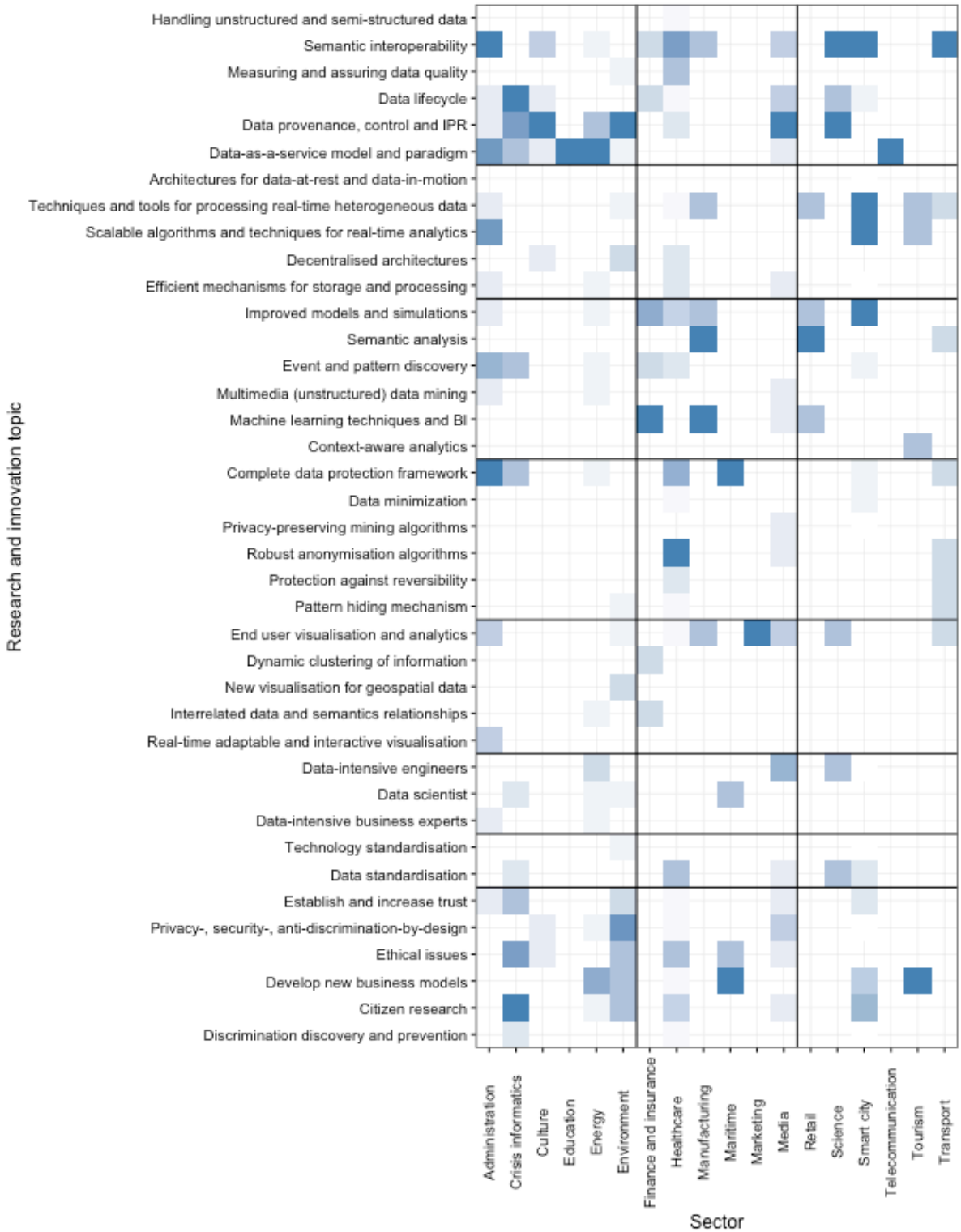


Figure 13. Impact of research in sectors, derived from BYTE analysis and the literature review. Relevance has been normalised by sector: darkest blue corresponds to the most relevant research in the sector.

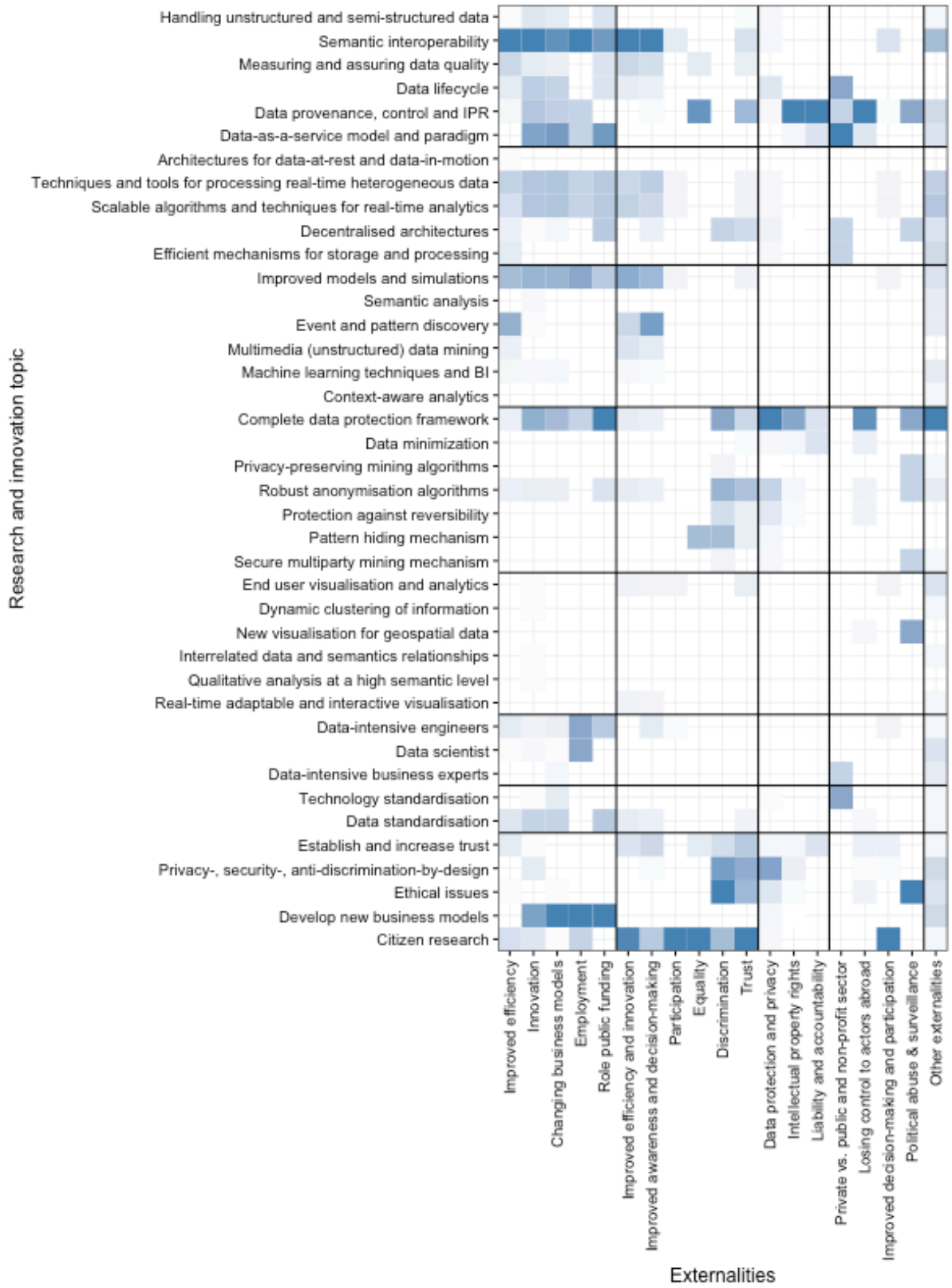


Figure 14. Impact of research in externalities, derived from BYTE analysis and the literature review. Relevance has been normalised by externality: darkest blue corresponds to the most relevant research in the externality.

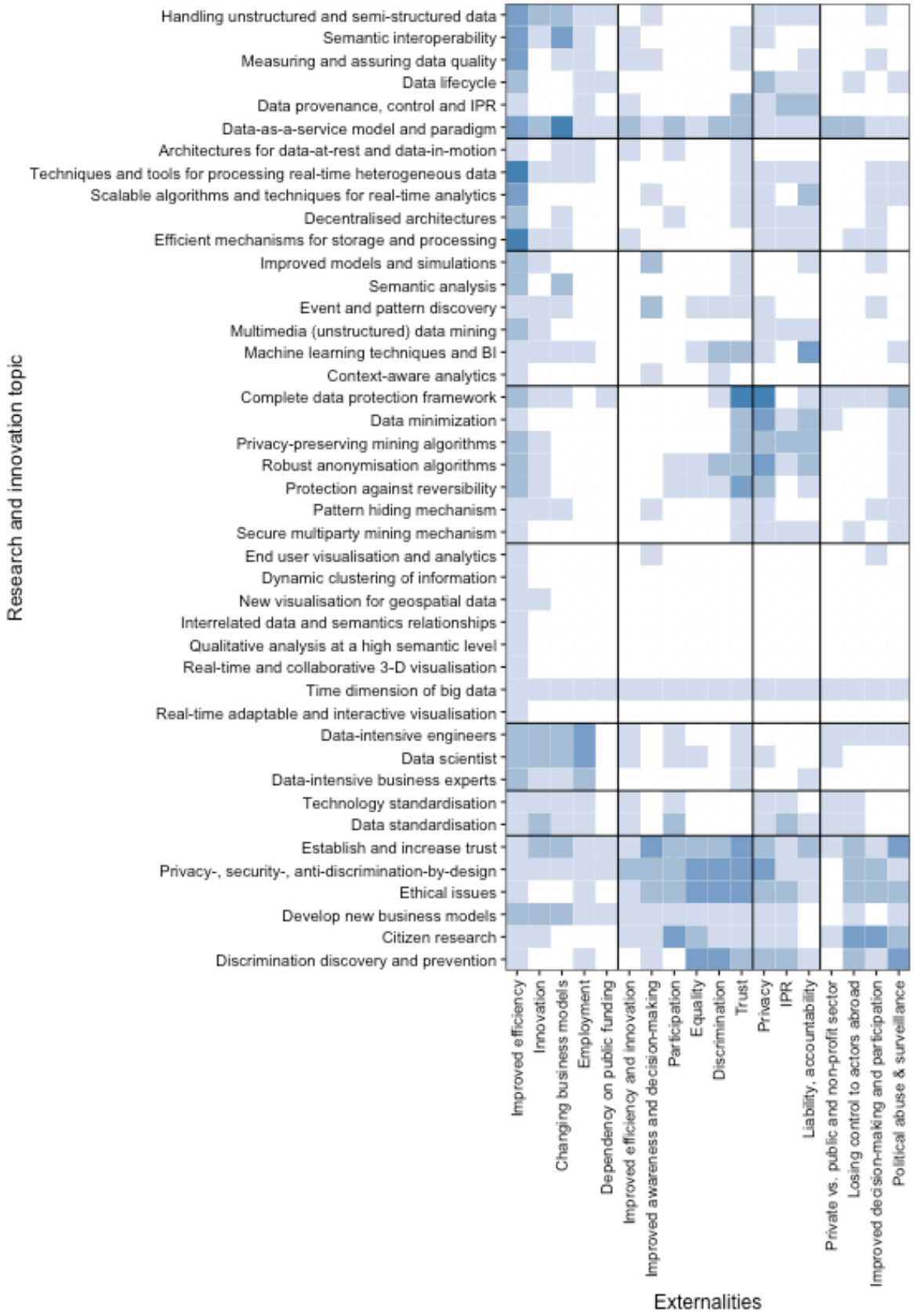


Figure 15. Impact of research topics in externalities, contributed by stakeholders and community members at the research roadmapping workshop. Priorities are coded as top (red), high (orange), medium (orange), low (green) and none (white).

- A common misconception of big data is to ignore "modelling, and instead rely on correlation rather than an understanding of causation" (Becker 2016) and that with enough data no models are needed (Anderson 2008). To address this issue, better modelling and simulations, and transparency about the data and the analysis to allow for a validation of the statistical significance of the results are recommended (Becker 2016). This includes taking into account design and sample biases.

Data protection:

- In industry, there is still a general fear of sharing data, which is partially compensated by the new value that is added by combining data. Possible solutions that were mentioned by stakeholders are the development of methods, possibly in the design phase and in the line of privacy-by-design, that can increase the trust in the protection of the data, and the development of mechanisms to encourage the emergence of more open business data, such as creating partnerships with public or research organisations that require or encourage open data publication.
- Enhanced cybersecurity captures the positive aspects of trust and privacy externalities.

Data visualisation was viewed as a lower-priority area:

- However, it was brought into attention that **better visualisations and user-friendly interfaces** might decrease the urgent **need of data skills** in the European market.

Non-technical priorities:

- **Data skills** for the general population will capture positive **employment** externalities, especially those connected with data-driven employment offerings and opportunities for economic growth through open data.
- Advances in **data standardisation** support communities and business **partnerships around data**.

3.3 ACTION PLAN

In this chapter, we present a prioritisation of research and innovation topics and timeline to tackle the societal externalities, develop the necessary skills and address the standardisation needs of big data in Europe, as well as general recommendations and best practices drawn from the BYTE research and contributions from the community. These prioritisations and recommendations will be further developed in the next chapter, for the three specific sectors (environment, healthcare and smart city) selected by the BYTE big data community as the ones on which to focus the community study and recommendations of the first year after project completion.

In the last part of the chapter, we provide a detailed list of actionable items for the 5 most relevant agendas that underpin multiple externalities and should command immediate attention and investment.

3.3.1 Research timeline

In this section, we present a timeframe to address the research and innovation topics, along with a prioritisation performed together with stakeholders and the community at the research roadmapping workshop. Topics have been categorised by a priority scheme as described in section 3.2.3: top (red), high (orange), medium (yellow) and low (green) priority. The timeframe spreads in detail over five years (2017-2021) and includes as well topics to be addressed in the mid- and long-term. Finally, and as presented in section 3.1.2, the roadmap visualises three different means by which these innovations can deliver an impact into the

different sectors and externalities: through **standardisation** (Figure 16), **societal impact** (Figure 17) and **skills development** (Figure 18), as well as by other means (Figure 18).

Figure 16. Timeline for research topics with impact on standardisation. Priorities are coded as top (red), high (orange), medium (orange), low (green) and none (white).

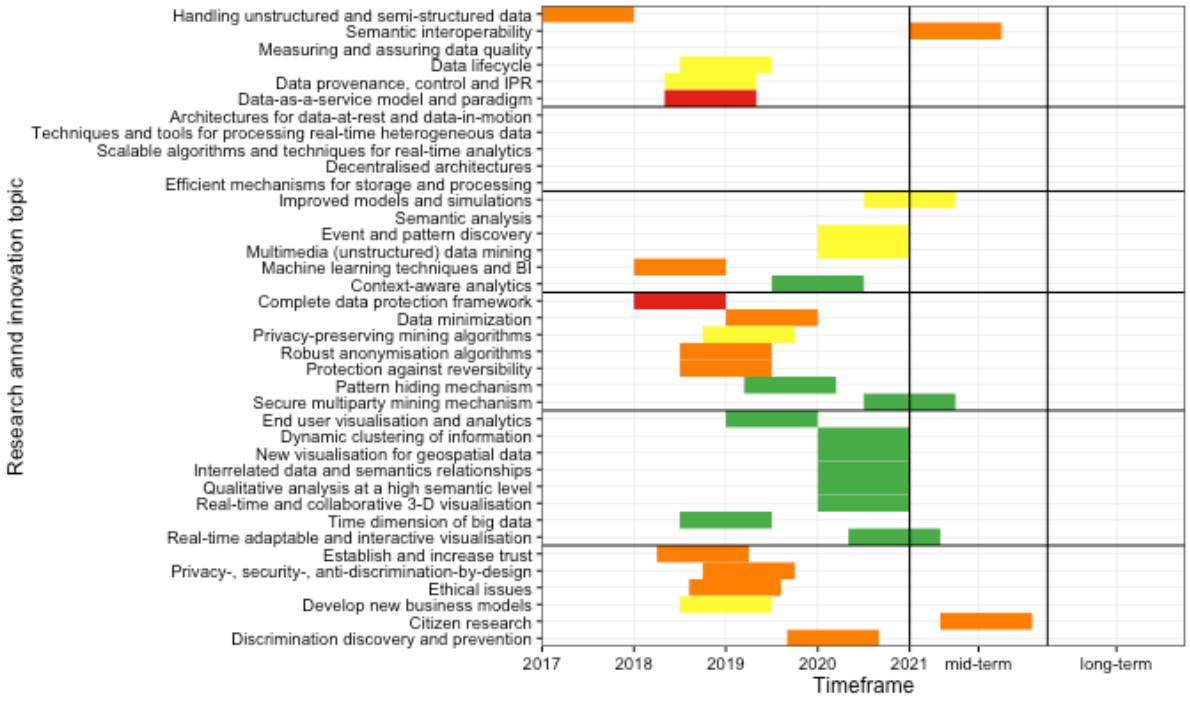


Figure 17. Timeline for research topics with societal impact. Priorities are coded as top (red), high (orange), medium (orange), low (green) and none (white).

Figure 18. Timeline for research topics with impact on skills development. Priorities are coded as top (red), high (orange), medium (orange), low (green) and none (white).

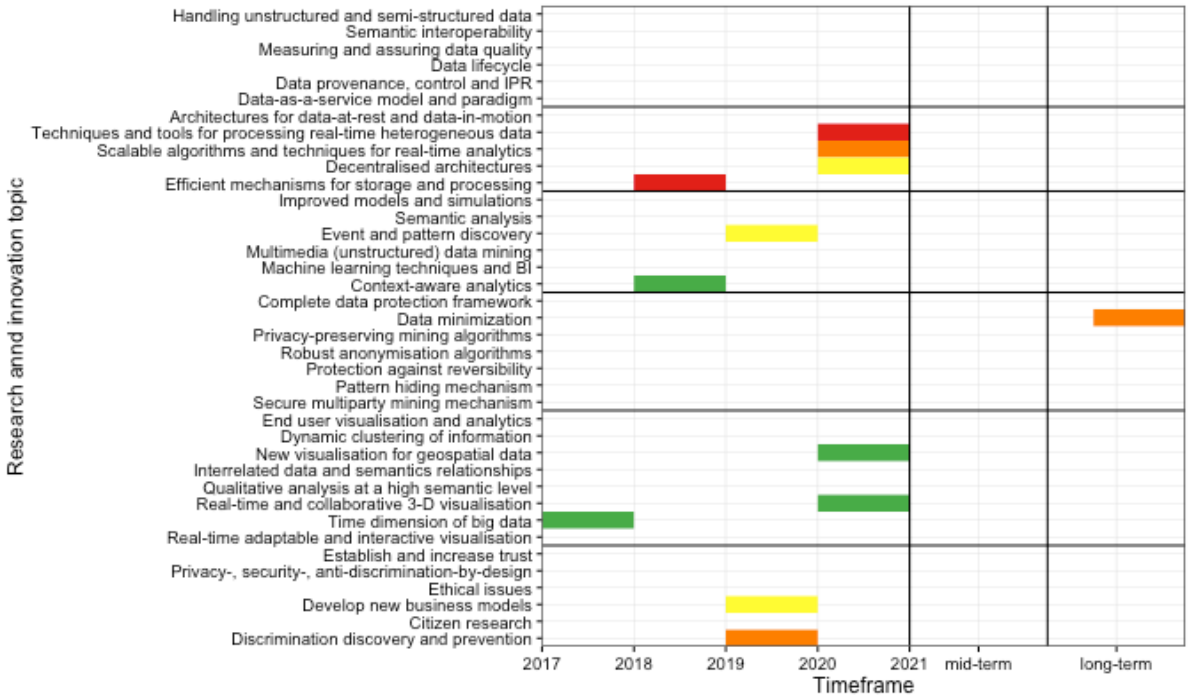
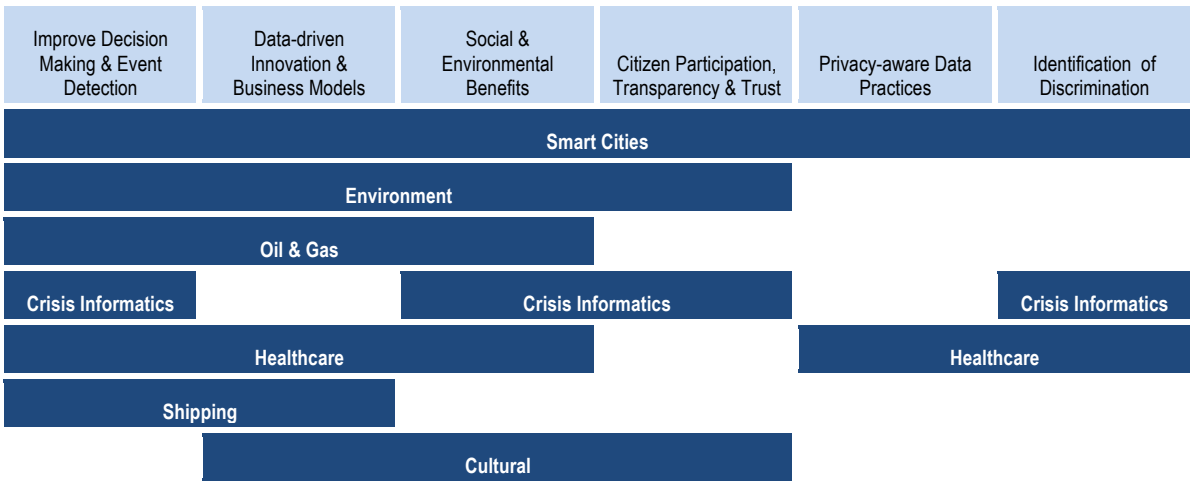


Figure 19. Timeline for research topics with other relevant impact. Priorities are coded as top (red), high (orange), medium (orange), low (green) and none (white).

We expect research and innovations in these topics to address the negative externalities and deliver positive social benefits. In this regard, BYTE has identified six areas where such positive benefit will be especially relevant. They are summarised in Table 9 along with their relevance to the BYTE case studies, and further described in *Big data for good* (Cuquet, Vega-Gorgojo, et al. 2016, 18-22) and in (Cuquet, Vega-Gorgojo, et al. 2017). These benefits are **data-driven innovations and business models**, the use of data analytics for large volumes of data to **improve event detection, situational awareness, and decision making** to e.g. allocate

resources efficiently, better **environmental protection and efficiency and direct social impact to citizens** through e.g. individual targeted services, the use of big data to enable **citizen participation and increase transparency and public trust** (this will require efforts to develop data skills among the general public), an increased attention paid to **privacy and data protection** by big data practitioners, and big data as a means to **identify and combat discrimination**.

Table 9. Mapping of the social benefits of big data against different sectors. Table reproduced from (Cuquet, Vega-Gorgojo, et al. 2016, 18).



3.3.2 Best practices

In order to capture these benefits, several best practices have been suggested by the BYTE project (Lammerant, De Hert and Vega Gorgojo, et al. 2015, 5, Cuquet, Vega-Gorgojo, et al. 2016, 23-24), which are further addressed in the policy part of this roadmap:

- **Public investments in infrastructures**
- **Funding programs for big data**
- **Public investments in open government data**
- **Persuade "big actors" to release some of their data**
- **Promote big data in education policies**
- **Look for interesting data sources**
- **Seek opportunities in new business models**
- **Create partnerships around data**

They involve public investments and funding programs to solve the scarcity of European big data infrastructures, promote research and innovation in big data, open more government data and persuade big private actors to release some of their data as well, so data partnerships can be built around them. New data sources and business models also need to be promoted. Interoperability has also been shown to be a key enabling factor. In addition, education policies have to address both the current scarcity of data scientists and engineers, but also the inclusion of data skills in general educational programs.

To address discrimination, equality and trust, privacy-by-design methods should be extended to anti-discrimination-by-design and analogous approaches, and transparency and new accountability frameworks need to be based both on legislation and on a data protection framework. Overall, policy makers, regulators and stakeholders all have an important role in updating legal frameworks, promoting big data practices and developing and incorporating tools into the big data design and practice that address societal concerns.

These best practices can also be followed to capture positive social benefits associated to social externalities. Furthermore, investment in the **interoperability of big data** is also a key recommended action (Lammerant, De Hert and Vega Gorgojo, et al. 2015, 6).

Another best practice to address negative social and ethical externalities regarding the risk of discrimination e.g. due to bias in the problem definition, data mining or training data is to use **auditing tools and extend privacy-by-design to anti-discrimination-by-design** (Lammerant, De Hert and Vega Gorgojo, et al. 2015, 6).

Regarding legal externalities, and besides the need to adapt regulations on a policy level, to address the non-scaling legal frameworks in the context of a high amount of interactions, it is recommended to substitute legal mechanisms based on individual transactions or individual control models with **aggregate or collective mechanisms and develop "by-design"-approaches** that translate legal objectives into technical requirements (Lammerant, De Hert and Vega Gorgojo, et al. 2015, 6-7). Another recommendation is to **develop standardised solutions and a toolbox of legal, organisational and technical means to fine-tune data-flows** (Lammerant, De Hert and Vega Gorgojo, et al. 2015, 7). Finally, **privacy-by-design** has to include not only a technical perspective but also legal and organisation safeguards to address the overall capabilities and risks of the systems (Lammerant, De Hert and Vega Gorgojo, et al. 2015, 7).

3.3.3 Recommendations

To provide an effective set of recommendations based on the mapping found above, we have performed a correspondence analysis (Greenacre, 1983) of the BYTE cases studies, literature research and workshop contributions (see Figures 15 and 16). This allows to identify the three most relevant dimensions with the actions needed to foster research with an impact on society, what is this impact on society, what are the opposing forces coming from both research and society that will present challenges and risks and the actors involved, which are taken from the classification of actors in BYTE case studies (Cunningham et al., 2016, p. 20-21) and their objectives (Cunningham et al., 2016, p. 26-30).

Dimension 1: Privacy and discrimination

Impact

This dimension is concerned with the protection of privacy and against discriminatory practices, as well as of intellectual property rights. It is also connected with the minimisation of political abuse and surveillance risks. It has also a secondary impact on the trust deposited by citizens in companies and public organisations, and aims to avoid losing control to actors abroad.

The concern over privacy, intellectual property rights and, to a lesser extent, discrimination was present among almost all case studies. The exception of the oil and gas case study should not be extrapolated to the whole energy sector, as it has been repeatedly seen that there is indeed a citizen concern on privacy issues, e.g. regarding smart meters (McDaniel & McLaughlin, 2009).

Challenges and risks

Changing business models, innovation and improved efficiency are commonly perceived as the main challenge to privacy. Despite of this, the BYTE case studies have found that big data has had the positive effect of an increased attention on privacy, and each sector provides examples on how to address it (Cuquet et al., 2017).

The advancement in research areas such as semantic interoperability, machine learning and real-time data processing poses the risk of opening a door to discriminatory practices and threats to privacy. The following actions should take into account this risk and be included in the research mentioned above.

Actors

- Citizens, concerned on their control of own data and their privacy.
- Policy-makers, interested in effective governance of big data for all Europeans, equal access and opportunities, and effectiveness of a balanced privacy regulation with due protection of individuals and appropriate use of data.
- Research institutions
- Companies, who are obliged to comply with laws and regulatory frameworks around privacy, data protection, non-discrimination and equal treatment
- Regulators (e.g. data protection authorities) and agencies responsible for oversight and enforcement of the law

Actions

- Regulate **research that analyses human data**, and establish a framework that allows companies to contribute robustly anonymised data to the scientific community. Quantify and mitigate risks of sharing datasets than can be later recombined and protect against reversibility.
- Address **bias in big data processes**, e.g. sample bias introduced by technical, economic or social factors or subjective bias from data labelling.
- Research on **legal informatics and algorithm accountability**.
- Adopt an overall risk-based approach to privacy and data protection, concerned with anonymisation but also with the full technical, legal and organisational safeguards. This includes the extension of privacy-by-design and other by-design approaches to cover all these safeguards.
    - This will require well-funded, -staffed and –resourced data protection authorities in order to ensure that the law is enforced and observed by those utilising big data

Dimension 2: Data accessibility for the digital economy

Impact

The second dimension deals with the development of new and changing business models, based on a new use of data or by specialisation in specific services, and the creation of partnerships around open data and precompetitive research. New and changing business models can have both a positive and a negative impact (Cuquet et al., 2017); the challenges they pose are outlined below.

This dimension is also connected with the relationships between private, public and non-profit sector, and between big businesses and SMEs. Tensions between private and public or non-profit organisations exist for example in the crisis informatics, culture and smart cities sectors. These are caused by lack of long-term commitment and dependency relations, inability to cooperate and future uncertainties, respectively (Cuquet et al., 2017).

A key priority to support these impacts is the improvement of data quality, open data and data discoverability.

Challenges and risks

- There is a risk of a **competitive disadvantage for SMEs [and start-ups]**, increased by the creation of a few dominant market players. Traditional non-digital services are also challenged to adapt to new business models.
- **Employment losses** for certain job categories, particularly if compounded by a failure for social security/public assistance systems to adapt to a changing 'big data' economy
- **Privatization** of essential utilities.
- There is a challenge to develop end user visualisation and analytics and foster citizen research in order to enhance **decision-making and participation** by the community, and reduce the negative impact of changing business models.

Actors

- SMEs/start-ups, implementing new business models.
- Large companies, co-opting market entry.
- Policy-makers, promoting equal-access to and opportunity in the market and supporting effective access to open data.
    - Regulators responsible for governing monopolies, business practices etc.
- Research institutions

Actions

- Continue the public **support to open data**. In particular, public funding should keep prioritising research that supports open data initiatives [and encourages public participation? Thus fostering community participation, etc.]. In academic research, tracking and recognition of data and infrastructure has to be improved. In this direction, **open dataset publication should be recognised** analogously to paper publication in journals for the purposes of performance evaluation of researchers and encouraging re-use and re-interpretation of datasets
- Develop **search engines for datasets**, with e.g. ranking algorithms analogous to the standard search engines, with the aim of making already existing data easily discoverable and usable. Develop best practices for data and technology rather than standardisation.
- Develop **data curation by demonstration** algorithms, analogous to programming by example or by demonstration.

Dimension 3: Participation and equality

Impact

The last dimension by its impact on citizen participation and equality, and has also a secondary impact on employment. A great potential of participation enabled by big data was ubiquitous in all case studies. Such big data effect on participation was twofold: directly from individual citizens, and also as a support to communities and civil society organisations. Citizen science initiatives and crowdsourcing has been successfully used in many scenarios related to the environment and smart cities sectors (Cuquet et al., 2017). To fully provide a societal benefit of big data, equality in data skills, access, infrastructure or language as to be promoted as well.

Challenges and risks

- Participation and engagement may be hampered by the fear of data abuse and privacy violations, lack of knowledge of digital/data tools and skills, inaccessibility of data presented (e.g. use of specialist/technical language).
- There is a risk of a large, unmet need of data scientists and engineers. This can be greatly mitigated by an increase in data skills in the general public and in key expert

positions. Data-intensive policy makers are an example of a skill that was identified to be of high priority.
- Open data, that supposedly benefits citizens in general, may increase the digital divide and produce social inequality, as data is only effectively open to a small elite. The divide affects gender inequality as well

Actors

- Citizens, in all their dimensions of autonomy and agency: control, context and transparency, learning, acquisition and input.
- Policy-makers, funding and promoting citizen science initiatives and supporting data-literacy programs.
- Media and journalism
- Research institutions
- Companies – for example use of big data/digital tools held and operated by private companies for targeted advertising (whether commercial or political), individualised pricing systems, etc. can have an effect on political participation, participation in social and public life/economic activities

Actions

- Citizen science initiatives are instrumental in involving citizens and increasing as well transparency and trust.
- Equip citizens with data skills. There is a need for specialists (data scientists and data-intensive engineers) that has been extensively identified. However, a **population-wide data literacy** is also equally important and has to be promoted through curriculum changes throughout all education stages. Moreover, stress has been put to integrate **ethics education** into the data science curricula.
- Promote **data journalism** to process, digest, and present the newly available open data to society.
- Ensure regulatory agencies are equipped with the expertise and resources necessary to enforce rules on privacy, data protection, anti-competitive practices, etc.

Conclusion

In this section, we have presented the three main dimensions in which big data research should focus to optimise the societal impact of big data: *privacy and discrimination*, *data accessibility for the digital economy*, and *participation and equality*. For each dimension, the main actions expected to deliver the desired impact are outlined. In Table 10, we present with more detail the relationship between the specific research and innovation priorities discussed in Section 3.2.1 with the actions of each dimension.

Table 10. Actions of each dimension and their related research and innovation priorities.

Dimension and actions	Research and innovation priorities
1. Privacy and discrimination	
Regulate research with human data	• Ethical issues • Protection against reversibility • Secure multiparty mining mechanism • Robust anonymisation algorithms • Data minimization • Privacy-preserving mining algorithms
Address bias in big data processes	• Ethical issues

Research on legal informatics and accountability	• Discrimination discovery and prevention
Adopt a risk-based approach to privacy and data protection	• Privacy-, security-, anti-discrimination-by-design • Data minimization • Complete data protection framework
2. Data accessibility for the digital economy	
Public support to open data	• Data-as-a-service model and paradigm
Data discoverability	• Data lifecycle • Technology standardisation
Data curation	• Data provenance, control and IPR
3. Participation and equality	
Citizen science initiatives	• Citizen research
Population-wide data literacy, including ethics education	• Ethical issues • Data-intensive engineers • Data scientists
Promote data journalism	• Establish and increase trust • Citizen research • Ethical issues

In the dimension of **privacy and discrimination**, ethical issues are especially relevant, e.g. as regards to the usage of human data in research. This includes regulations, but also development of standards that enable public and private organisations to contribute anonymised data for research purposes while minimising the risk of backfiring and research to mitigate the risks of de-anonymisation via recombination of datasets.

As mentioned before, bias is also a relevant ethical issue that needs further research. The implicit assumption of big data objectivity should be dismissed, and account explicitly for the potential sample bias introduced by technical, economic or social factors.

In conjunction to the challenge of bias in data, more research is also needed within the area of discrimination discovery and prevention and on legal informatics and accountability to account for algorithmic bias.

An overall risk-based approach to privacy and data protection requires research into anonymisation but also the development of privacy-by-design, anti-discrimination-by-design and analogous by-design approaches to cover the full technical, legal and organisational safeguards.

To promote **data accessibility for the digital economy**, the research effort should be concentrated in the area of data management. Opening data with an origin in industry is still an issue that can be addressed by technical means, and not only via legal and policy action. Public funding should keep prioritising processes that support open data initiatives. In research, it is recommended to recognise open dataset publication analogously to how paper publications are recognised for the purpose of performance evaluation of researchers and research institutions.

To facilitate the discoverability of already existing data, there is a need to develop search engines for datasets that include quality ranking, similar to how websites are ranked. This would have the double positive effects of surfacing data and driving data owners to publish better datasets, as discussed in the data lifecycle research priority. In addition, Data curation by demonstration, in analogy to programming by example or by demonstration, would also for the distribution and scalability of the system.

Finally, the dimension of **participation and equality** is chiefly influenced by citizen research. Three actions are proposed to this end. First, the promotion and extension of citizen science programmes and practices, both in cooperation with traditional research initiatives or as stand-alone projects. Second, the development of population-wide data literacy to minimise the risk of exclusion and of digital divide. To this aim, data skills and ethical considerations are recommended to be included in the curricula of all education levels and in particular of high-school education. Third, development of an independent data journalism that processes, digests and presents newly available open data to citizens is crucial for those that do not have the skills or time to make use of all such data. This action depends directly on the non-technical priority of establish and increase trust, and also on the area of data visualisation.

4 ROADMAP IMPLEMENTATION AND COMMUNITY ACTIONS

Aside from the broad recommendations and timeline to address the research topics presented above, the present roadmap also foresees an annual deeper study of selected sectors to be taken up initially by the BYTE project partners and community members, and by the BYTE community alone after project completion. Each year, a group of three sectors will thus be addressed in detail to produce special recommendations and actions. The BYTE big data community will gather and consolidate feedback from NGO, IGO, academy and other civil society experts about good practice in these sectors. The results will then be fed to industry and policy makers, and especially to the BDVA and other relevant networks and projects as outlined in the updated Interim strategy and charter for the big data community (Bigagli, et al. 2016).

The goal is that the community is able to present a deeper discussion on what and where are the gaps and challenges each sector faces, and recommend good practices and specific research and policy needs to cover these gaps.

For the first year, five sectors from the BYTE case studies were preselected based on their impact, relevant audience in the BDVA and active involvement by the community. These sectors were culture, energy, environment, healthcare and smart city. Of them, environment, healthcare and smart city were selected by the BYTE partners, BYTE advisory board and founding members of the BYTE big data community as the first three sectors to be further studied.

In the first part of this section, we address what are the specific research needs for these three sectors. These results are to be taken up and further analysed in the upcoming BYTE community workshop and then fed to the relevant channels.

In addition to these three vertical roadmaps, we also present four horizontal roadmaps that summarize the recommended actions in regards to privacy aware access control for big data, big data impact on society, big data education and big data analytics strategy.

The new version of the *Final sustainability plan for the big data community* presents in its Section 6 the future actions and timeline with which the BYTE big data community will continue the development and update of the present roadmap, and input its recommended actions into the Big Data Value Association and other networks (Bigagli et al., 2017, pp. 29-32). Here, in the last subsection we summarise such action plan.

4.1 ROADMAPS' IMPACT ON THE ENVIRONMENT SECTOR

4.1.1 Policy actions

From a policy point of view, the environment sector will be impacted by investments on technologies and infrastructures and public/private partnership:

- Fostering the **development of data standards** will help the development of such partnerships by allowing the exchange of data between a network of actors.
- **Investing on data storage and processing facilities along with encouraging the development of techniques such as data mining** will help store the relevant datasets and mine the necessary knowledge to model the current situation and identify the adequate tools to improve this situation.
- **Partnering with global platforms** will help understanding environmental issues at a global level and potentially address this issue globally. It could also help monitoring the footprint of users' activities and influence these activities.

- Eventually, **investments on sustainable computing paradigms**, such as green computing could help reduce the impact of the data transition the environment.

On the whole, the environmental issue is a long term one. Yet, we emphasized in our roadmap that some actions could be taken quickly and start making the big data transition a tool to answer this issue.

4.1.2 Research priorities

An overview of the externalities associated with the environment sector is given in (Cuquet, et al. 2016, 9-10). Figure 20outlines the research priorities relevant to the environment sector and a timeframe for their development. These are described below, with their specific impact in this sector.

Figure 20. Research and innovation topics with an expected impact on the environment sector.

Data management:

R-DM-03. Measuring and assuring data quality.

R-DM-05. Research into **data provenance, control and IPR** to minimise threats to intellectual property rights (including scholars' rights and contributions). This includes scalable data access mechanisms. It has to contribute to fix, on a technical level, the current lack of norms for data storage, processing and use.

R-DM-06. Data-as-a-service model and paradigm to exploit new opportunities for economic growth (new products and services based on open access to big data). New models should be encouraged that diminish inequalities to data access between big data players and the rest.

Data processing:

R-DPROC-02. Techniques and tools for processing real-time heterogeneous data that help gather public insight by identifying environmental trends and statistics.

R-DPROC-04. Open **decentralised architectures** that diminish storage costs and decrease the dependency on external data sources, platforms and services.

Data protection:

R-DPROT-06. Pattern hiding mechanism to avoid discriminatory practices and targeted advertising (as a result of profiling and tracking private data).

Data visualisation:

R-DV-01. End user visualisation and analytics need to ensure that manipulation of visual representations of data is avoided. This may require to formulate 3D and 4D ethics (R-SO-03).

R-DV-03. New visualisation for geospatial data to help understand and manage environmental data and enable data-driven policy-making. This has repercussions to other sectors as well that will benefit from easy access to such data reports.

Non-technical priorities:

R-SO-01. Establish and increase trust via better transparency and accountability of the public sector.

R-SO-02. The increasing awareness about privacy violations and ethical issues of big data can be met by developments in **privacy-by-design, security-by-design, anti-discrimination-by-design** frameworks. Such developments will decrease the public reluctance to provide information (and especially personal data), the threats to data protection and personal privacy, and contribute to overcome the reduced innovation due to restrictive legislation

R-SO-03. Investigate **ethical issues** around "sabotage" data practices, e.g. in social media fraud profiles willingly misinforming and possibly creating false data that affect the overall picture.

R-SO-04. Develop **new business models** with closer linkages between research and innovation to capture opportunities for economic growth based on open access to big data. Examples put forward by the BYTE case study were the use of sea data for fishing purposes and weather data in the tourism industry. These new models may contribute to diminish the dominance of big market players.

R-SO-05. The environment sector shows great potential to encourage **citizen research**, which may take the form of crowd-computing, pervasive-computing, crowd-sourcing or independent research using open data and tools. To enable it, such tools have to be developed. This would increase citizen participation, produce safe and environment- friendly operations, deliver better models, measures and test about preparedness and resilience of communities, as well as of human behaviour under crisis, and generally enhance quality of life. Furthermore, a strong participation of the public sector can help make data and services from the environment sector to become public goods available to all.

This sector also has a strong requirement of data scientists (R-SK-02) to cover new data-driven employment offerings that will result from the challenge of traditional non-digital services, such as traditional weather forecasting. Furthermore, technology (R-ST-01) and data standards (R-ST-02) need to be developed to enhance data-driven R&D. Finally, skill development should also be aimed at diminishing inequalities to data access and the data divide.

4.2 ROADMAPS' IMPACT ON THE HEALTHCARE SECTOR

4.2.1 Policy actions

As the environment sector, the healthcare sector will be impacted by investments on technologies and infrastructures and public/private partnership. The time targets of this impact depends on the time targets of the recommendation we made in the roadmap:

- There exist health data standards[40]. Yet, they could be improved or replaced by new standards which would ensure data quality, patient safety and the adoption of new tools such as digital clinical records. Such standards are necessary to allow the advancement of clinical research (Richesson and Krischer 2007).
- **Investing on data storage and processing facilities along with encouraging the development of techniques such as data mining** will help advancing clinical, and more widely, medical knowledge. Investing on educating data scientists and integrating data analysis into mainstream curricula will also help benefit from the data transition.
- **Partnering with global platforms** will also help doing so as exemplified by agreements between national health systems and data companies.
- Health data are very sensitive. **Auditing security mechanisms** to protect them, raising citizens' awareness and **opening a large scale debate** while investing in technologies such as privacy enhancing technologies is necessary to allow people to trust the data transition in the health sector (Goldberg and Ian 2003).

4.2.2 Research priorities

An overview of the externalities associated with the healthcare sector is given in (Cuquet, et al. 2016, 10-11). Figure 21outlines the research priorities relevant to the healthcare sector and a timeframe for their development (Cuquet, Fensel and Bigagli 2017). These are described below, with their specific impact in this sector.

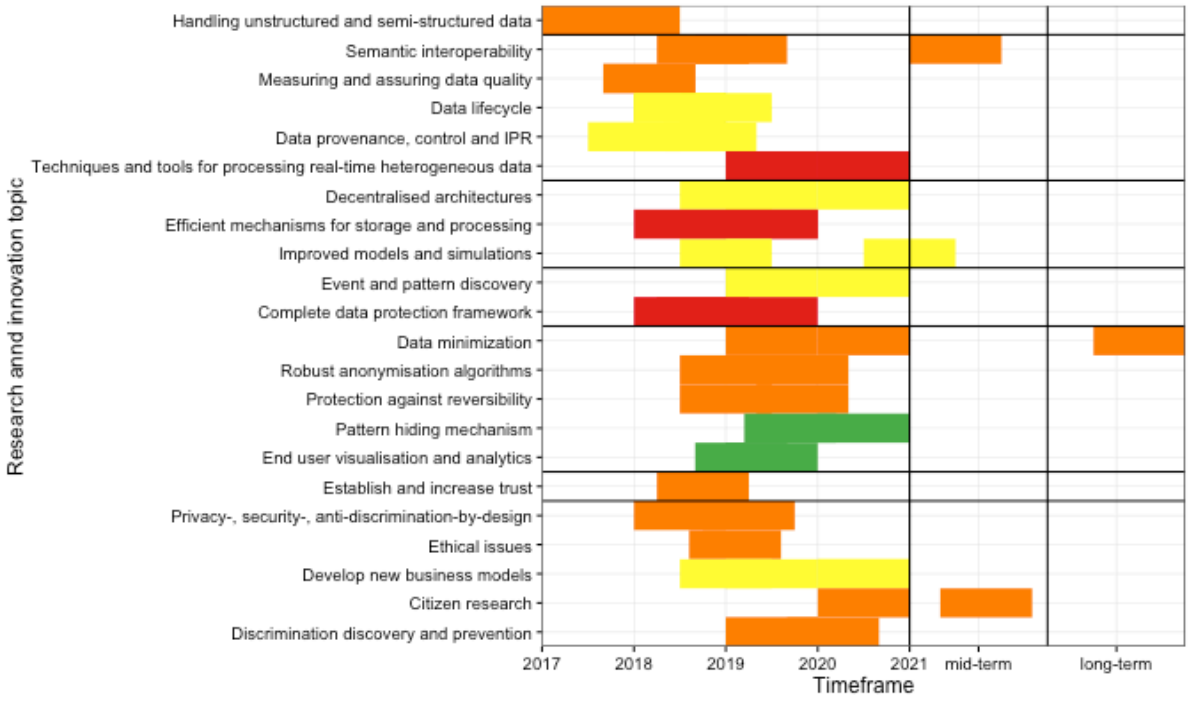


Figure 21. Research and innovation topics with an expected impact on the healthcare sector.

Data management

R-DM-01. Research and innovations in **handling unstructured and semi-structured data** to develop easy-to-use reporting tools that include semantic annotations, so extra and repetitive work is avoided. This can assist in providing context to avoid incorrect interpretations.

[40] http://www.hl7.org/implement/standards/

R-DM-02. Typically, separately generated pools of data in the healthcare sector have remained unconnected. Due to the high heterogeneity in data source, **semantic interoperability** innovations are needed to enable better healthcare services through advanced integration of heterogeneous health data, data sharing and analysis. Easy-to-use reporting tools need to be developed. They should include semantic annotations that do not add extra work. These interoperability challenges are often gaps in the market for innovative business models and the development of tools that achieve commercial viability for innovators.

R-DM-03. Only a small percentage of data is documented with low quality. Better services could be delivered by **measuring and assuring data quality**. An example is clinical decision support applications, which rely on data integration (R-DM-02) and a very high data quality so physicians can actually rely on them. Innovations in the area of data quality offer opportunities for economic growth, mainly through community building around data partnerships and sharing information across sectors.

R-DM-04. The whole **data management lifecycle** in healthcare offers opportunities for innovative business models and economic growth, as described above in R-DM-02 and R-DM-03.

R-DM-05. Like in the environment sector, more research into **data provenance, control and IPR** is needed to minimise threats to intellectual property rights (including scholars' rights and contributions), with the same focus as mentioned above in Section 4.3.1.

Data processing

R-DPROC-02. Techniques and tools for processing real-time heterogeneous data, e.g. in the area of stream data mining and analysis, can allow to make use of user-generated content from blogs, forums and social media to track disease outbreaks, side effects of drugs, etc.

R-DPROC-04. Decentralised architectures will benefit from the development of cryptographic mechanisms applicable to cloud and big data, such as attribute-based encryption, which diminish the threats to data protection and personal privacy.

R-DPROC-05. Efficient mechanisms for storage and processing that can efficiently cope with data quality and heterogeneity issues at a large scale will have a high impact in knowledge-driven sectors such as the healthcare sector.

Data analysis

R-DA-01. Improved models and simulations of clinical operations and complex analytics to provide new insights about the effectiveness of treatments will deliver better health services through data sharing and analysis.

R-DA-03. Event and pattern discovery includes social media analysis to identify e.g. clusters of posts that can assist to track unplanned events (like accidents) and improve emergency response. It includes also research into dynamic social processes such as the spread of diseases and how big data can contribute to it.

Data protection

R-DPROT-01. Data security and privacy issues hinder data exchange in healthcare, and need to be addressed by advances in the **complete data protection framework**. The storage, processing, access and protection and big data depends the development of a legal framework or guidelines that is backed by a parallel technical one. As an example, and patients may want to opt-in or out of incidental findings related to their personal health data.

R-DPROT-02. Data minimization should be more widely adopted to reduce the risk of private data misuse, especially when sharing with third parties without consent.

R-DPROT-04. Researchers are already aware of anonymity needs. However, **robust anonymisation algorithms** are needed to reduce the risk of privacy threats even with anonymised data and with data mining. Promising lines of research are recent developments like k-anonymity.

R-DPROT-05. In relation also to the previous item, better **protection against reversibility** is needed. As the potential uses of health data open up with emerging technologies, developments in this area are relevant to overcome public reluctance to provide information (especially personal data) so data can be shared among different public and private organisations without risk of deanonymisation..

R-DPROT-06. Pattern hiding mechanisms that can deal with the fingerprints that certain diseases and genetic disorders combinations can provide, especially when those that are rare.

Data visualisation

R-DV-01. End user visualisation and analytics is relevant in all sectors, also in healthcare and the related life sciences and pharmaceutical sectors. Developments in this area should support prediction and classification processes, as well as assist in explorative analysis.

Non-technical priorities

R-SO-01. Establish and increase trust (transparency, efficacy, manageability and acceptability).

R-SO-02. Developments in **privacy-by-design, security-by-design, anti-discrimination-by-design** as described in the previous environment sector section.

R-SO-03. Ethical issues around patient involvement and privacy. Private data misuse, especially sharing with third parties without consent, needs to be addressed too. In the case of analysis for emergency response and crisis situations that rely on social media data and other data dependent on the unequal access to technology due to e.g. economic or social factors, it has to be further explored how deal with the possible data bias, so to ensure that data is used to provide a benefit to all society and avoid an increase in inequalities due to different data access or production. Moreover, and as mentioned above (R-DPROT-01), specific to the healthcare sector is the ethical questions raised by incidental findings, also when using data-driven tools.

R-SO-04. Develop new business models around pre-competitive partnerships for data curation to open opportunities for economic growth through community.

R-SO-05. Citizen research based on self-monitoring and self-sensing (also called "quantified self").

From a policy point of view, the availability of big amounts of data will enable politicians to have more information about different situations in the health sector and thus a better understanding that may lead to improve their decision-making and increases the investments in healthcare. Regarding standardisation needs, in this sector there is a lack of standardised health data that affects the analytics usage and could be addressed via standard electronic health records and common data models and ontologies.

4.3 ROADMAP'S IMPACT ON THE SMART CITY SECTOR

4.3.1 Policy actions

Smart cities are both a political and a technological challenge. They will strongly influence other sectors, such as the environment and health. Specifically, smart cities will be impacted by the following recommendations of the policy roadmap:

- **Investing in data storage and processing facilities along with encouraging the development of techniques such as data mining** will help local governments and institution understand the territory they govern and identify the right course of actions.
- **Partnering with data companies** will also help local policymakers retrieve useful datasets - on mobility for instance - while investing in technologies such as pattern mining will help them model the population they govern and potentially influence its activities. This could involve monitoring resources consumptions and promoting green modes of transportation for instance.
- **Investing on turning local governments as platforms** could help simplify the interaction between citizens and their local governments. Coupled with **improving digital literacy and an open data strategy**, it could help involving citizens.
- **Building a safe and trustworthy data infrastructure** is necessary to guaranty the acceptation of smart cities. This could rely on opening a democratic debate on data, promoting privacy enhancing technologies.
- Eventually, **distributing the benefits of local sharing economy** while addressing issues such as tax collection is necessary to ensure the fairness of the data transition.

4.3.2 Research priorities

An overview of the externalities associated with the environment sector is given in (Cuquet, et al. 2016, 11-12). Figure 22 outlines the research priorities relevant to the environment sector and a timeframe for their development. These are described below, with their specific impact in this sector.

Data management

R-DM-02. Semantic interoperability for urban multimodal transportation, sensor data, social media data and user-generated data from e.g. citizens' smartphones, etc. This will deliver more targeted services for citizens, impact the cost-effectiveness of services and optimise utilities, increase citizen participation, gather public insight, enhance the energy efficiency of the city, foster innovation from open data, and create new economic and innovative opportunities via community building across different sectors.

R-DM-04. Data management lifecycle that opens data to foster innovation from open and government data. The public investment into a data infrastructure for the city and its subsequent opening should allow more value to be created by the engaged citizens and start-ups, which will naturally be drawn to such cities.

Figure 22. Research and innovation topics with an expected impact on the smart city sector.

Data processing

R-DPROC-01, R-DPROC-02 and R-DPROC-03. Innovative usage of **architectures for data-at-rest and data-in-motion**, **techniques and tools for processing real-time heterogeneous data** and **scalable algorithms and techniques for real-time analytics** to allow optimisation of utilities through data analytics will contribute to an efficiency increase in the city. This will contribute to the impact mentioned in R-DM-02.

R-DPROC-05. Efficient mechanisms for storage and processing to contribute to the R-DM-04 topic.

Data analysis

R-DA-01. Improved models and simulations based on newly integrated heterogeneous data sources (R-DM-02).

R-DA-03. Event and pattern discovery, centred e.g. on anomaly detection using traffic sensor data to get accurate traffic estimates and allow for a more efficient organisation. Additionally, and as mentioned in the healthcare sector above, social media post clustering can identify unplanned events (like accidents) and improve emergency response.

Data protection

R-DPROT-01 and R-DPROT-02. The new sources of data in the smart city sector create new ways of possible data misuse. A new **data protection framework**, together with **data minimisation**, should aim at protecting individuals first, rather than their data.

R-DPROT-03. To promote **privacy-preserving algorithms**, open source machine learning algorithms should be encouraged in order to increase trust, in the same way as cryptography algorithms are put under public scrutiny.

Non-technical priorities

R-SO-01. Establish and increase trust through opening analysis algorithms (see also R-DPROT-03) and allow citizens to easily visualise and understand what big data reveals of themselves.

R-SO-02. As mentioned above in the data protection framework, **privacy-by-design, security-by-design, and anti-discrimination-by-design** frameworks should be developed that put the protection of the citizen first.

R-SO-04. Develop new business models around a publicly funded and promoted open data infrastructure.

R-SO-05. A smart city represents an optimal playground for community engagement on local political issues, data collection and analysis in which **citizen research** and crowd-sourced applications can be pursued.

The smart city sector is in a big demand of data-intensive engineers, more than of data-scientists. Regarding standards, both data and technology standards (R-ST-01, R-ST-02) are much needed in the smart city, and will be probably be set de facto by bigger cities possibly outside Europe, as the big size of one market (city) determines the standards for the rest. To contribute to this standardisation, European cities should group together to create a significant market pull. This is especially relevant in order to avoid a competitive disadvantage of newer business and local SMEs in front of large dominant player from abroad.

4.4 PRIVACY AWARE ACCESS CONTROL FOR BIG DATA: A RESEARCH ROADMAP FOR EUROPE

Privacy has been identified as a relevant externality in all studied sectors. Privacy issues are indeed a relevant issue that concerns the public (Donovan et al., 2015). Aside from policy and legal recommendations outlined in the present document and in Lammerant, De Hert, Vega Gorgojo et al. (2015), there are also technical challenges that have to be addressed, such as fine-grained access controls (Freitas and Curry, 2016) and digital rights (Qin and Atluri 2003). With the raise of big data, the major security challenges are now in non-relational data stores (Colombo and Ferrari, 2015a; Strohbach et al. 2016). Privacy-by-design and related approaches that extend the concept to e.g. anti-discrimination are generally seen as a good solution to the privacy and trust challenge (Lammerant, De Hert and Vega Gorgojo, et al. 2015, 6-7; Domingue et al. 2016). Taking a risk-based approach is also important. Even if privacy risks might be minimal, it has to be taken into account that they affect a very large population with highly varying privacy expectations (Metcalf, Keller and Boyd 2016, 17). It is also important that, although further research is still needed in these areas, many interesting approaches already exist but are still not well known in industry (NESSI, 2012, p. 25). In this section, we present the main actions needed for a privacy-aware access control (PAAC) for big data. The recommendations are based on an already existing roadmap for PAAC (Colombo and Ferrari, 2015a) and aligned to their societal implications in Europe based on the BYTE case studies and vision. Figure 23 outlines the research priorities relevant to the privacy aware access control roadmap and a timeframe for their development. The recommended actions are shown below.

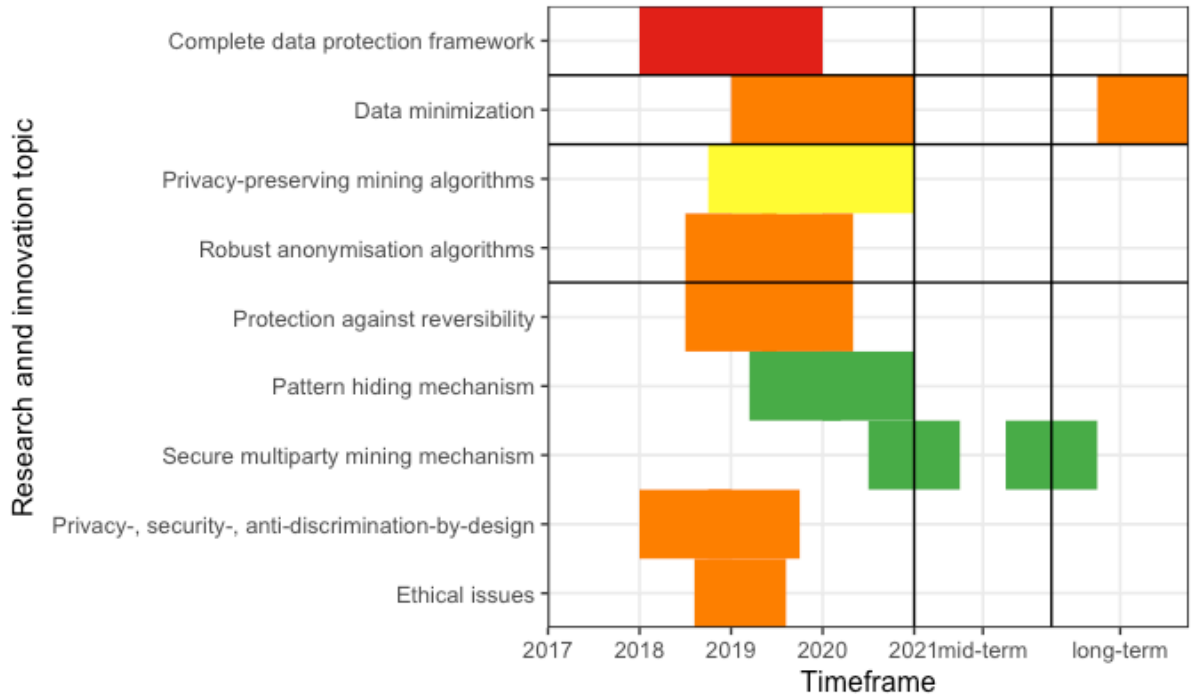


Figure 23. Research and innovation topics of the privacy aware access control roadmap.

Actions

- **R-DPROT-01.** Develop granular access controls to allow sharing data on a fine-grained level.
- **R-DPROT-01, R-DPROT-07.** Develop mechanisms for access to sensible and identifiable data only with aggregated results. This has been already proposed for relational databases (Colombo and Ferrari, 2015b) and should be extended to NoSQL datastores as well.
- **R-DPROT-02, R-DPROT-03, R-SO-02.** Develop privacy-by-design and related by-design approaches that incorporate not only technical but also legal and organisational safeguards. Extend the use of auditing tools.
- **R-DPROT-04, R-DPROT-05, R-DPROT-06, R-SO-03.** Quantify and minimise the risks.
- **R-DPROT-03.** Support the dissemination and transfer of already existing solutions from research to industry.

4.5 BIG DATA IMPACT ON SOCIETY: A RESEARCH ROADMAP FOR EUROPE

The whole research roadmap has been focused on the impact of big data on society. The three most important dimensions of this impact have been identified by means of correspondence analysis (Greenacre, 1983) as *privacy and discrimination*, *data accessibility for the digital economy* and *participation and equality*. Section 3.3.3 provides a summary of the societal impact, challenges, risks and actors associated with these three dimensions. Figure 24 outlines the research priorities relevant to the society roadmap and a timeframe for their development. Below, we compile the recommended actions.

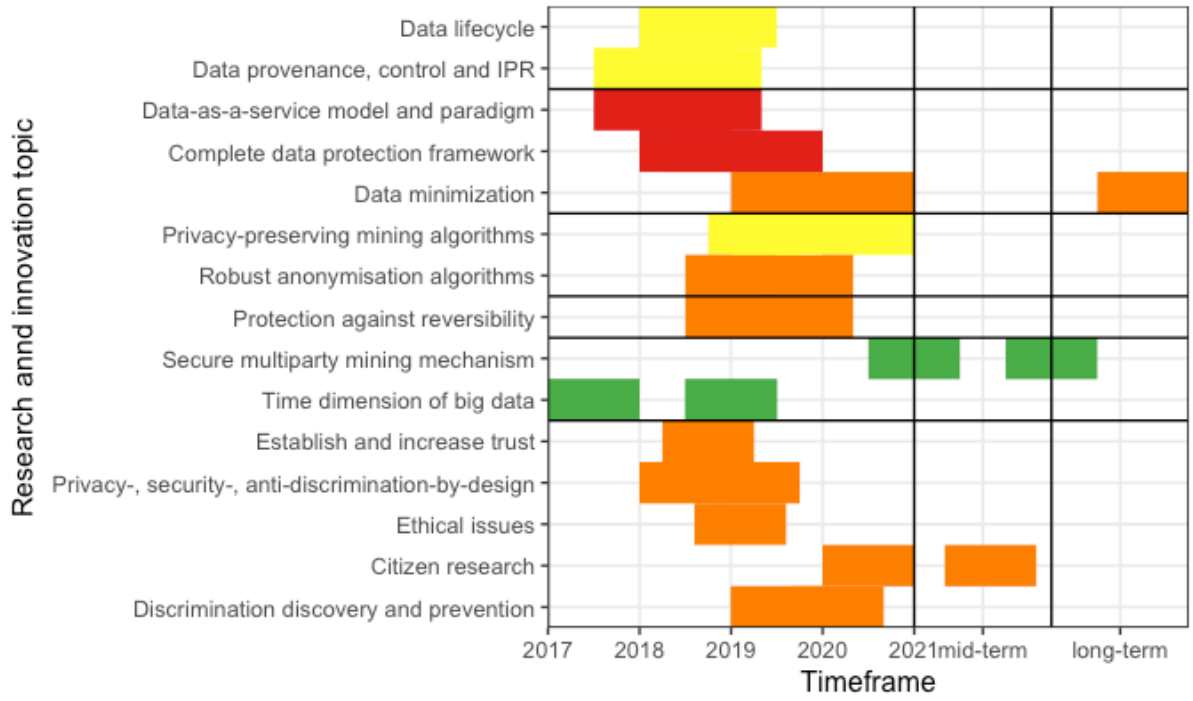


Figure 24. Research and innovation topics of the society roadmap.

Actions

Privacy and discrimination

- **R-SO-03, R-DPROT-02, R-DPROT-03, R-DPROT-04, R-DPROT-05, R-DPROT-07.** Regulate **research that analyses human data**, and establish a framework that allows companies to contribute robustly anonymised data to the scientific community. Quantify and mitigate risks of sharing datasets than can be later recombined and protect against reversibility.
- **R-SO-03.** Address **bias in big data processes**, e.g. sample bias introduced by technical, economic or social factors or subjective bias from data labelling.
- **R-SO-06.** Research on **legal informatics and algorithm accountability**.
- **R-SO-02, R-DPROT-01, R-DPROT-02.** Adopt an overall **risk-based approach** to privacy and data protection, concerned with anonymisation but also with the full technical, legal and organisational safeguards. This includes the extension of privacy-by-design and other by-design approaches to cover all these safeguards.

Data accessibility for the digital economy

- **R-DM-06.** Continue the public **support to open data**. In particular, public funding should keep prioritising research that supports open data initiatives. In academic research, tracking and recognition of data and infrastructure has to be improved. In this direction, **open dataset publication should be recognised** analogously to paper publication in journals for the purposes of performance evaluation of researchers.
- **R-DM-04, R-ST-01.** Develop **search engines for datasets**, with e.g. ranking algorithms analogous to the standard search engines, with the aim of making already existing data easily discoverable and usable. Develop best practices for data and technology rather than standardisation.
- **R-DM-05.** Develop **data curation by demonstration** algorithms, analogous to programming by example or by demonstration.

Participation and equality

- **R-SO-05.** Citizen science initiatives are instrumental in involving citizens and increasing as well transparency and trust.
- **R-SO-03, R-SK-01, R-SK-02.** Equip citizens with data skills. There is a need for specialists (data scientists and data-intensive engineers) that has been extensively identified. However, a **population-wide data literacy** is also equally important and has to be promoted through curriculum changes throughout all education stages. Moreover, stress has been put to integrate **ethics education** into the data science curricula.
- **R-SO-01, R-SO-05, R-DV-07, R-SO-03.** Promote **data journalism** to process, digest, and present the newly available open data to society.

4.6 BIG DATA EDUCATION: A RESEARCH ROADMAP FOR EUROPE

As mentioned in Section 3.2.2, the need for educated people equipped with the right data skills has been extensively identified (see e.g. Manyika, et al. 2011, NESSI 2012, Mattmann 2013, e-skills uk 2013, Curry, et al. 2014, 32, Berger, et al. 2014, 50-51). The three required skill profiles that are commonly identified are *deep analytical talents / data scientists*, *data-savvy managers and analysts / data-intensive business experts* and *supporting technology personnel / data-intensive engineers* (Manyika, et al. 2011; Big Data Value Association, 2016, p. 33). The BYTE case studies and research roadmapping workshop has confirmed this need, and pointed out that also *policy- and decision-makers* need to be data-savvy as well. The urgent need of data skills in the European market could be addressed by better visualisations and user-friendly interfaces, as pointed out by several members of the BYTE big data community. The development of industry-focused and high-education curricula is already under way. For example, the European Data Science Academy project[41] is addressing this challenge and has released a report that evaluates the skills gap and how to close it (Mack, Tarrant and Dadzie 2016) and proposed a data science curriculum (Phethean and Simperl, 2016). Many new Bachelor and Master degrees are now available in the European Higher Education Area. Some very recent examples are the Universitat Oberta de Catalunya new degrees of BSc in Data Science and MSc in Data Science. However, developing population-wide data skills is still a challenge that is not being fully tackled and that poses the risk of increasing the digital divide, as well as gender inequalities (Roberts, 2012; Lammerant, De Hert and Vega Gorgojo, et al. 2015, 30). This issue may be addressed by developing data competences in secondary education in Europe. Both this secondary education and the high education curricula should also take into account ethics around data practices (Metcalf, Keller and Boyd 2016, 13). Finally, data journalism could have an important role in processing, digesting, and presenting the newly available open data to society and in the dissemination of these skills. Figure 25 outlines the research priorities relevant to the education roadmap and a timeframe for their development. The recommended actions are shown below.

[41] https://edsa-project.eu

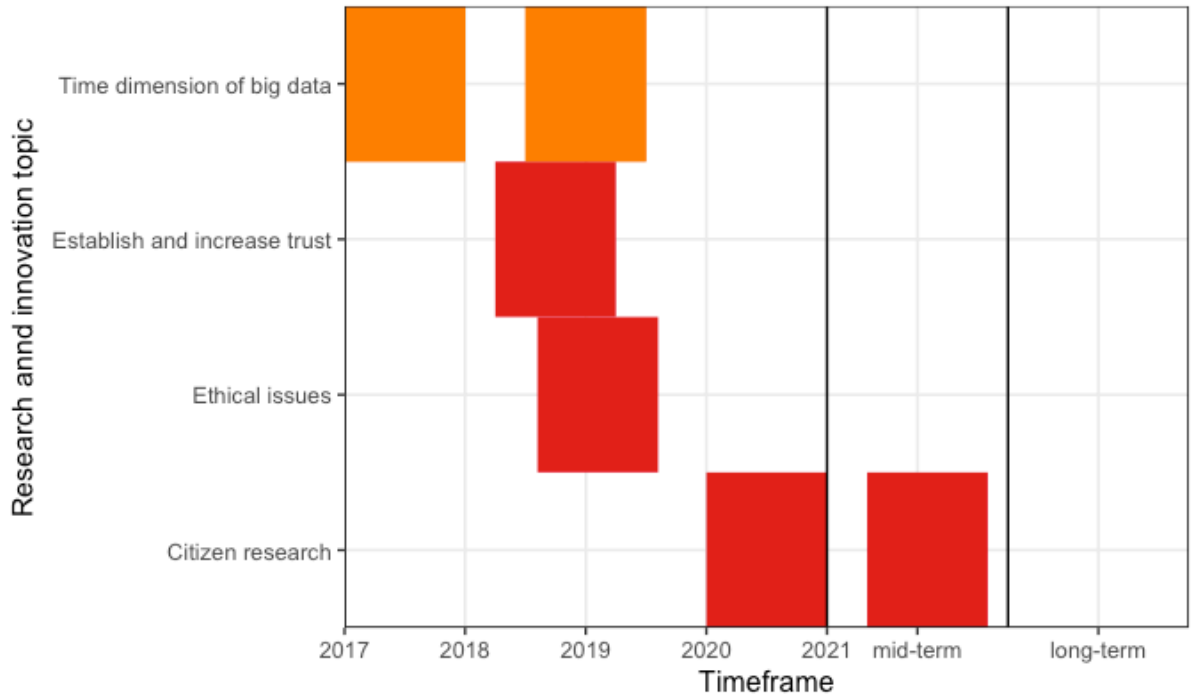


Figure 25. Research and innovation topics of the education roadmap.

Actions

- **R-SK-01, R-SK-02, R-SK-03.** Continue the development of high education data science curricula.
- **R-SO-01, R-SO-03, R-SO-05.** Develop **curricula for secondary education** that promote data competences.
- **R-SK-01, R-SK-02, R-SK-03, R-SO-03.** Integrate **ethics education** in both secondary education and high education curricula.
- **R-SO-01, R-SO-05, R-DV-07, R-SO-03.** Support **data journalism** as a means to disseminate skills and ethical issues.

4.7 BIG DATA ANALYTICS STRATEGY: A RESEARCH ROADMAP FOR EUROPE

The priorities for a big data analytics strategy have been correctly identified by the Big Data Value Association (2017, p. 29-30). Europe has strong competitive advantages in this area, which is expected to improve efficiency and add value to any sector, as well as having the role to provide better access to big data to the wider public (Big Data Value Association, 2017, p. 29). On of such areas where Europe has an opportunity is to lead multilingual sentiment analysis, and in general multilingual analysis tools. However, there are some issues and challenges that are sometimes overlooked and that have been found to be of the highest priority.

In general, most analytics models would extremely benefit by methods to correct sample bias and the representativeness of data (Becker 2016). This affects also the data collection and quality assessment processes. In a related note, there is also an urgent need of validated methodologies and standards behind big data analytics and especially in those that are used as a base or justification to take decisions, so that decision-makers can easily identify and correctly assess that the recommendations provided by data and models are trustworthy and apply to the problem at hand. In such predictive and prescriptive models, and also in event and pattern discovery, there is also the need to further investigate and differentiate between correlation and causation. In this direction, an evidence-driven, bottom-up approach has been put forward by (Brodie 2015) to first deduce correlations from evidence (eg using data from

economic phenomena) and then develop means to estimate their correctness and completeness, such as the probabilistic likelihood that correlations are causal within error bounds. Figure 26 outlines the research priorities relevant to the analytics roadmap and a timeframe for their development. The recommended actions are shown below.

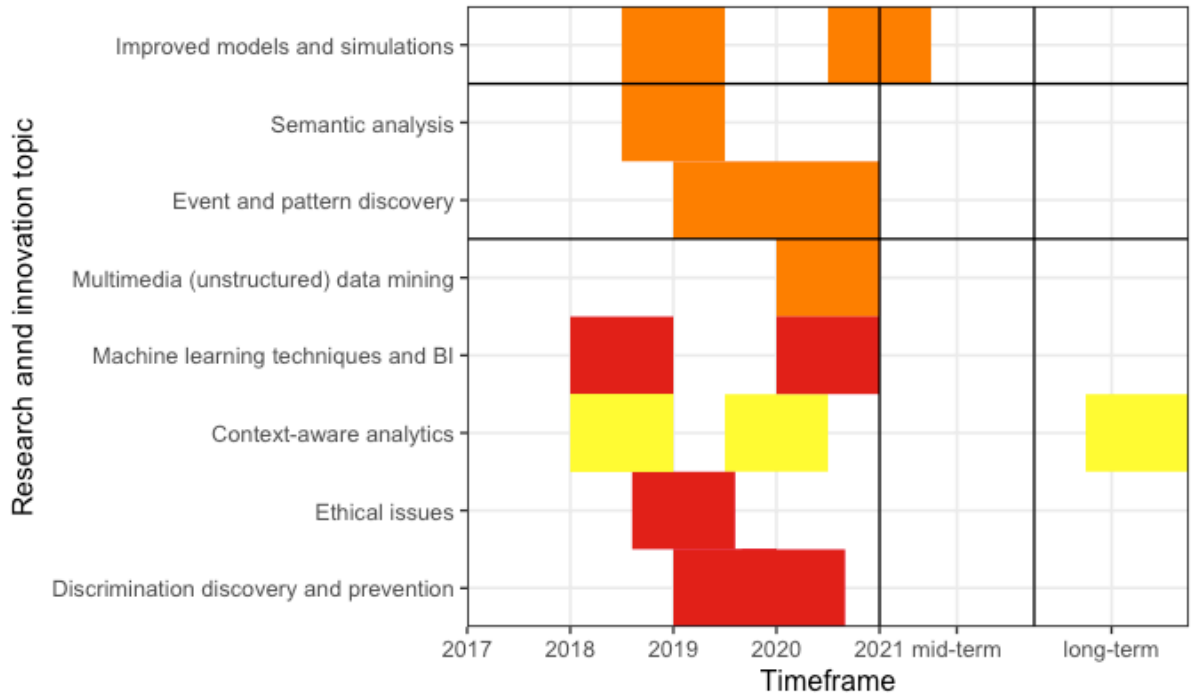


Figure 26. Research and innovation topics of the analytics roadmap.

Actions

- **R-DA-05, R-SO-03, R-SO-06.** Asses **sample biases and data representativeness** in predictive and prescriptive analytics, and differentiate between correlation and causation.
- **R-DA-05.** Differentiate between **correlation and causation** in predictive and prescriptive analytics.
- **R-DA-01, R-DA-02, R-DA-04.** Develop multilingual sentiment analysis, and in general **multilingual analytics** tools.
- **R-SK-03, R-ST-01. Validate methodologies and standards** behind the analytics on whose results decisions are to be taken, and that are easily identifiable and understandable by decision-makers.

4.8 FUTURE ACTIONS AND TIMETABLE

In this subsection, we summarise the action plan of the BYTE big data community (BBDC) to further develop and update, and input its recommendations into the Big Data Association and other relevant networks. Such plan is discussed in more detail in the *Final sustainability plan for the big data community* (Bigagli et al., 2017, pp. 29-32). This work is part of the three objectives of the BBDC to continue the BYTE research on societal externalities of big data, how to make the best of them, engage NGOs, NPOs, CSOs, third sector, local governments, tech-transfer organizations, and Input and feedback to EC, Member States, BDVA, membership and related networks.

The BBDC focuses each year on three areas that extend the previous BYTE research and update the roadmap. It began in 2016 with the environment (Cuquet, Fensel and Bigagli, 2017), healthcare and smart city sectors. The BBDC focus areas for 2017 are energy, transport and

trusted artificial intelligence in smart industry. The planned research and consultation on current big data issues, good practices and touch points with standards and methodology will take place between August and December 2017. In 21-23 November 2017, a workshop on these focus areas will be collocated with the European Big Data Value Forum in Versailles, France, jointly organised by the Big Data Value Association and the European Commission. This will include also a report on the focus area with the intention to provide input to the BDVA SRIA, EC work programme and other relevant networks. This format will be repeated for 2018 and 2019 with the selection of three new focus areas per year, the corresponding research and consultation and their input into the correspondent networks and organisations. In 2020, it is planned to consider the merging of the BBDC into the Big Data Value Association, which will presumably take over the tasks.

5 CONCLUSION

We have presented one of the main outcomes of the BYTE project: a roadmap for big data in Europe. This roadmap is divided in two parts: the policy roadmap provides a set of recommendations to build a policy framework which will ensure that Europe can make the most out of the data transition. The research roadmap focuses on identifying the research priorities at stake.

Both the research and policy roadmaps present different perspectives that when combined outline a clear set of priorities for European policy and research from the perspective of economic, legal, social, ethical and political externalities. When read together the policy and research dimensions, actions and priorities achieve a number of commonalities. First, all three of the dimensions for both research and policy map quite well against one another. There are significant overlap in actions and priorities within the *data governance dimension* identified in the policy roadmap and the *privacy and discrimination dimension* considered in the research roadmap. Both include a consideration of data governance, data protection, privacy legislation and privacy-related technical and methodological innovations. Similarly, the *data accessibility dimension* and the *infrastructure investment dimensions* of the respective roadmaps cover issues such as data standards, technology standards, new business models, new service offerings and the centrality of government as a key actor to achieve the actions and priorities identified. Finally, the *social good dimension* and *participation and equality dimension* also have key areas of overlap including citizen involvement in research, policy making and data driven activities, the need for greater data literacy and the need to increase trust and transparency through policy and technological or methodological innovations.

Nevertheless, in each case, there are issues specific to the perspectives being offered. The policy roadmap is more concerned with policies that will solve economic and political challenges alongside legal, social and ethical issues. This focus makes sense as technology innovations are often neutral until they become embedded in wider social systems. Thus, the policy roadmap focuses on the digital single market, cross-border data arrangements, government services, public investment and resource management. In contrast, the research roadmap focuses more on technological capabilities, with a foregrounding of capabilities to better ensure legal, social and ethical protections. These include privacy-preserving algorithms, privacy-by-design, data security mechanisms, data lifecycles, citizen science and data journalism.

When mapped against one another, however, the policy and research roadmaps highlight a set of overlapping concerns that are crucial to enable European actors to achieve a greater share of the big data market. Table 11 maps these dimensions actions and priorities to identify overlaps, including:

- Data protection
- Privacy
- Transparency in data processes
- Support for open data
- Data standards
- Data accessibility (incl. sharing agreements)

Table 11: Research and policy dimensions, actions and priorities

Research dimension	Action	Research priorities	Policy priorities	Action	Policy dimension
Privacy and discrimination	Regulate research with human data Address bias in big data processes Research on legal informatics and accountability Adopt a risk-based approach to **privacy and data protection**	Ethical issues Protection against reversibility Secure multiparty mining mechanism Robust anonymisation algorithms Data minimization **Privacy-preserving mining algorithms** Discrimination discovery and prevention **Privacy-, security-, anti-discrimination-by-design** Complete protection framework	Single market Cross-border data agreements **Open data** **Privacy management**	**Facilitating data transfers** Foster innovation **Foster transparency**	**Data governance**
Data accessibility for the digital economy	Public support to **open data** **Data discoverability** Data curation	Data-as-a-service model and paradigm Data lifecycle **Standardisation**	**Promote data standards** Invest in innovative sectors	Data formats and standards Public investment in data infrastructures	**Infrastructure emergence**

		Data provenance, control and IPR	Invest in infrastructure Digitalise disrupted sectors Digitise government services	Digital business models	
Participation and equality	Citizen science initiatives Population-wide **data literacy,** including ethics education Promote data journalism	**Citizen research** **Ethical issues** Data-intensive engineers Data scientists **Establish and increase trust**	**Digital literacy** Data security **Benefits and risks to the digital economy** Make better use of data for public service provision	**Citizen participation** Resources management Sharing benefits of data transition	**Social good**

- Digital literacy
- Citizen participation and engagement
- Increasing trust

These issues both align with and depart from the external research and policy roadmaps examined within this project and this report. The BYTE roadmaps are more policy focused than the European technical roadmaps and focus largely on how technical innovations can aid in achieving non-technical, societal requirements. For example, technical innovations like protections against reversibility or privacy-preserving algorithms can be used to meet legal and social privacy and ethical requirements. In roadmaps like the BDV SIRA ((Freitas and Curry 2016), these are subsumed as "non-technical priorities", despite some attention to automated data governance. In contrast, while other roadmaps that focus on political or economic issues often focus on policy, the BYTE roadmaps include a consideration of how technical or procedural innovations can contribute to societal goals. The BYTE roadmaps also thread an awareness of the importance of stakeholder collaboration through their recommendations, recognising that in addition to policy-makers, other actors also need to be involved. Specifically, in order to achieve digital literacy or sufficient infrastructure, other actors, including research institutions, industry, educational institutions and citizen groups need to be involved.

Thus, the BYTE research and policy roadmap is oriented towards a range of different stakeholders who will be mobilised to consider their contributions to these policy and research priorities, including policy-makers, civil society organisations, institutions, industry bodies and academics. The BYTE Big Data Community will work directly with established big data organisations in Europe, but particularly the BDVA, to implement this roadmap and keep it updated as part of European research and policy setting practices and processes. This process is described in greater detail in Deliverable 7.1.2: *Final strategy and charter for the big data community*.

6 BIBLIOGRAPHY

Akerkar, Rajendra, et al. “Understanding and mapping big data.” D1.1 BYTE Project, 31 March 2015.

Anderson, Chris. “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.” *Wired*, 23 June 2008.

Becker, Tilman, Anja Jentzsch, and Walter Palmetshofer. *Cross-sectorial roadmap consolidation.* D2.5 BIG Project, 21 November 2014.

Berger, Helmut, et al. *Conquering Data in Austria. Technologie-Roadmap für das Programm IKT der Zukunft: Daten durchdringen - Intelligente Systeme.* Vienna: Bundesministerium für Verkehr, Innovation und Technologie, 2014.

Big Data Value Association. “Big Data Value Strategic Research and Innovation Agenda.” Version 2. January 2016.

Big Data Value Association. “Big Data Value Strategic Research and Innovation Agenda.” Version 3. January 2017.

Bigagli, Lorenzo, et al. *Interim strategy and charter for the big data community.* D7.1.1 BYTE Project, 1 March 2016.

Bigagli, Lorenzo, Magnusson, J., Cuquet, M, Fensel, A. *Final sustainability plan for the big data community*. D7.2.2 BYTE Project, 2017.

Brodie, Michael L. “Understanding Data Science: An Emerging Discipline for Data Intensive Discovery.” *RCDL*. 2015.

Cranor, Lorrie, Tal Rabin, Vitaly Shmatikov, Salil Vadhan, and Daniel Weitzner. *Toward a Privacy Research Roadmap for the Computing Community.* White paper, Washington D.C.: Computing Community Consortium committee of the Computing Research Association, 2015.

Colombo, Pietro, Elena Ferrari. Privacy Aware Access Control for Big Data: a Research Roadmap. *Big Data Research* 1:4 (2015a): 145-154.

Colombo, Pietro, Elena Ferrari. Efficient Enforcement of Action-Aware Purpose-Based Access Control within Relational Database Management Systems. *IEEE Transactions on Knowledge and Data Engineering* 27:8 (2015b): 2134-2147.

Cunningham, Scott W., et al. *Tackling the Externalities of the Vision.* D5.2 BYTE Project, 15 March 2016.

Cuquet, Martí, and Anna Fensel. “Big data impact on society: a research roadmap for Europe.” *arXiv*, October 2016: 1610.06766.

Cuquet, Martí, Guillermo Vega-Gorgojo, Hans Lammerant, Rachel Finn, and Umair ul Hassan. “Big data for good.” D9.5 BYTE Project, 30 September 2016.

Cuquet, Martí, Guillermo Vega-Gorgojo, Hans Lammerant, Rachel Finn, and Umair ul Hassan. “Societal impacts of big data: challenges and opportunities in Europe” arXiv:1704.03361. 2017.

Cuquet, Martí, Anna Fensel, and Lorenzo Bigagli. “A European research roadmap for optimizing societal impact of big data on environment and energy efficiency”, *2017 IEEE Global Internet of Things Summit (GIoTS) Proceedings*, 2017.

Curry, Edward, et al. *Final Version of Technical White Paper.* D2.2.2 BIG Project, 28 02 2014.

Domingue, J., N. Lasierra, A. Fensel, T. van Kasteren, M. Strohbach, A. Thalhammer, Big Data Analysis, in: J.M. Cavanillas, E. Curry, W. Wahlster (Eds.), New Horizons a Data-Driven Econ., Springer International Publishing, 2016: pp. 63–86. doi:10.1007/978-3-319-21569-3_5.

Donovan, Anna, et al. "Report on legal, economic, social, ethical and political issues." D2.1 BYTE Project, 30 September 2014.

Donovan, Anna, R. Finn, K. Wadhwa. "Report on public perceptions and social impacts relevant to big data." D2.2 BYTE Project, 20 July 2015.

e-skills uk. "Big Data Analytics. An assessment of demand for labour and skills, 2012-2017." January 2013.

English, M., Auer, S., & Domingue, J. (2016, May). Block chain technologies & the semantic web: A framework for symbiotic development. In *Computer Science Conference for University of Bonn Students, J. Lehmann, H. Thakkar, L. Halilaj, and R. Asmat, Eds* (pp. 47-61).

Fensel, A., Toma, I., García, J. M., Stavrakantonakis, I., & Fensel, D. (2014). Enabling customers engagement and collaboration for small and medium-sized enterprises in ubiquitous multi-channel ecosystems. *Computers in Industry*, *65*(5), 891-904.

Forston, Lucy, et al. "Galaxy Zoo: Morphological Classification and Citizen Science." *Advances in Machine Learning and Data Mining for Astronomy*, March 2012.

Freitas, A., E. Curry, Big Data Curation, in: J.M. Cavanillas, E. Curry, W. Wahlster (Eds.), New Horizons a Data-Driven Econ., Springer International Publishing, 2016: pp. 87–118. doi:10.1007/978-3-319-21569-3_6.

Groth, Paul, Andrew Gibson, and Jan Velterop. "The anatomy of a nanopublication." *Information Services and Use* 30, no. 1-2 (2010): 51-56.

Gruber, T. "Towards principles for the design of ontologies used for knowledge sharing." In *Formal ontology in conceptual analysis and knowledge representation*. Knowledge Systems Laboratory, Stanford University, 1993.

Guarino, N., R. Poli, T.R. Gruber, Toward Principles for the Design of Ontologies Used for Knowledge Sharing, in: Form. Ontol. Concept. Anal. Knowl. Represent., Kluwer Academic Publishers, 1993.

Gutiérrez-Rubí, Antoni. "Gobernar el 'Open Data'." *eldiario.es*, 12 September 2016.

Hajirahimova, M. S., & Aliyeva, A. S. (2015). Big Data strategies of the world countries: Национальный Суперкомпьютерный Форум (НСКФ-2015), Россия, Переславль-Залесский, 24-27 ноябрь, 2015, 10, 11.
URL:http://2015.nscf.ru/TesisAll/8_Integraciya_visokoyrovnevix_resyrsov/09_402_AliyevaAS.pdf

Huang, (2016). *Big Data Initiatives in China: Opportunities and Challenges*. Keynote talk at 2016 IEEE International Conference on Research, Innovation and Vision for the Future on Computing and Communication Technologies (IEEE-RIVF'16), 7-9 November 2016, Hanoi, Vietnam. URL: http://rivf2016.tlu.edu.vn/Portals/11/Keynote2_RIVF2016.pdf.

International Energy Agency. *Energy Technology Roadmaps: a guide to development and implementation.* Paris: International Energy Agency, 2014.

Kamara, Seny, and Kristin Lauter. "Cryptographic cloud storage." *Proceedings of Financial Cryptography: Workshop on Real-Life Cryptographic Protocols and Standardization.* Springer, 2010. 136-149.

Kärle, E., Fensel, A., Toma, I., & Fensel, D. (2016). Why are there more hotels in Tyrol than in austria? analyzing schema. org usage in the hotel domain. In *Information and Communication Technologies in Tourism 2016* (pp. 99-112). Springer, Cham.

Khatib, Firas, et al. "Crystal structure of a monomeric retroviral protease solved by protein folding game players." *Nature Structural & molecular biology* 18 (2011): 1175-1177.

Lammerant, Hans, et al. *Horizontal analysis of positive and negative societal externalities.* D4.1 BYTE Project, 31 August 2015.

Lammerant, Hans, Paul De Hert, Guillermo Vega Gorgojo, and Erik Stensrud. *Evaluating and addressing positive and negative societal externalities.* D4.2 BYTE Project, 31 December 2015.

Le Novère, Nicolas, and Camille Laibe. "MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology." *BMC Systems Biology* 1 (2007): 58.

Lunzer, Aran, and Kasper Hornbæk. "Subjunctive interfaces: Extending applications to support parallel setup, viewing and control of alternative scenarios." *ACM Transactions on Computer-Human Interaction (TOCHI).* ACM, 2008. 17.

Lyko, K., M. Nitzschke, A.-C. Ngonga Ngomo, Big Data Acquisition, in: J.M. Cavanillas, E. Curry, W. Wahlster (Eds.), New Horizons a Data-Driven Econ., Springer International Publishing, 2016: pp. 39–61. doi:10.1007/978-3-319-21569-3_4.

Mack, Leonard, David Tarrant, and Aba-Sah Dadzie. "Study Evaluation Report 2." D1.4 EDSA Project, 29 July 2016.

Manyika, James, et al. "Big data: The next frontier for innovation, competition, and productivity." McKinsey Global Institute, June 2011.

Mattmann, Chris A. "A vision for data science." *Nature* 493 (January 2013): 473-475.

McDaniel, Patrick, Stephen McLaughlin, Security and Privacy Challenges in the Smart Grid, IEEE Security & Privacy 7:3 (2009).

Metcalf, Jacob, Emily F. Keller, and Danah Boyd. "Perspectives on Big Data, Ethics and Society." The Council for Big Data, Ethics and Society, 2016.

Munoz, C., Smith, M., & Patil, D. (2016). Big data: A report on algorithmic systems, opportunity, and civil rights. *Executive Office of the President. The White House*.

NESSI. "Big Data: A New World of Opportunities." NESSI White Paper, December 2012.

Papachristos, George, Scott W. Cunningham, and Claudia Werker. *The BYTE Vision.* D5.1 BYTE Project, 29 February 2016.

Peroni, Silvio, Alexander Dutton, Tanya Gray, and David Shotton. "Setting our bibliographic references free: towards open citation data." *Journal of Documentation* 71, no. 2 (2015): 253-277.

Phaal, Robert, Clare J.P. Farrukh, and David R. Probert. "Technology roadmapping—A planning framework for evolution and revolution." *Technological Forecasting & Social Change* 71 (2004): 5–26.

Phethean, C., E. Simperl. “Data Science Curricula 2”. D2.2 EDSA Project, July 2016.

Qiao, X., Xue, S., Chen, J., & Fensel, A. (2015). A lightweight convergent personal mobile service delivery approach based on phone book. *International Journal of Communication Systems*, *28*(1), 49-70.

Qin, Li, and Vijayalakshmi Atluri. “Concept-level access control for the Semantic Web.” *XMLSEC '03 Proceedings of the 2003 ACM workshop on XML security.* ACM, 2003. 94-103.

Roberts, Tony. “The problem with Open Data.” *ComputerWeekly.com*, 2012.

Silver, David, et al. “Mastering the game of Go with deep neural networks and tree search”, *Nature* 529 (2016): 484-489.

Şimşek, U., Fensel, A., Zafeiropoulos, A., Fotopoulou, E., Liapis, P., Bouras, T., Saenz, F. T., & Gómez, A. F. S. (2016, September). A semantic approach towards implementing energy efficient lifestyles through behavioural change. In *Proceedings of the 12th International Conference on Semantic Systems* (pp. 173-176). ACM.

Sharma, S. (2016). Expanded cloud plumes hiding Big Data ecosystem. *Future Generation Computer Systems*, *59*, 63-92.

Strohbach, M., J. Daubert, H. Ravkin, M. Lischka, Big Data Storage, in: J.M. Cavanillas, E. Curry, W. Wahlster (Eds.), New Horizons a Data-Driven Econ., Springer International Publishing, 2016: pp. 119–141. doi:10.1007/978-3-319-21569-3_7.

Sweeney, Latanya. “k-anonimity: a model for protecting privacy.” *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10, no. 5 (2002): 557-570.

Thanos, Constantino. “A Vision for Open Cyber-Scholarly Infrastructures.” *Publications* 4, no. 2 (2016): 13.

United States. Executive Office of the President, & Podesta, J. (2014). *Big data: Seizing opportunities, preserving values*. White House, Executive Office of the President.

Vega-Gorgojo, Guillermo, et al. *Case study reports on positive and negative externalities.* D3.2 BYTE Project, 5 June 2015.

Vega-Gorgojo, Guillermo, Grunde Løvoll, Thomas Mestl, Anna Donovan, and Rachel Finn. “Case study methodology.” D3.1 BYTE Project, 30 September 2014.

Wan, Z, et al. “A game theoretic framework for analyzing re-identification risk.” *PLoS One* 10, no. 3 (March 2015): e0120592.

APPENDIX 1: THE BIG DATA POLICY ROADMAP - SURVEY

1. What would you call "policy" in big data?
   A Laws
   B Charters
   C Standards and norms
   D Other

2. Policies influence a lot of actors involved in big data. Which actor do you consider influenced by policy?

   A Small enterprises
   B Large enterprises
   C Policymakers
   D Citizens
   E Other

3. What is the influence of a big data policy on these actors and their activities?*

   A It helps business development
   B It prevents business development
   C It helps existing businesses to move through their digital transition
   D It protects existing businesses
   E It may foster social good
   F It may help to achieve a better transparency
   G Other

4. Do you want to elaborate on the influence of big data policy?

5. In your opinion, what are the three main objectives of a big data policy?

6. Please rate from 1 star to 4 stars the objectives for big data policy to address
   Data governance (open data, public/private partnerships…)
   Social good
   Business/Ecosystem development
   Do you see a more important objective?

7. Let's talk a bit more about [FIRST OBJECTIVE][42]
   Who should be responsible for achieving this?
   How would you assess the completion of [FIRST OBJECTIVE]?
   When should [FIRST OBJECTIVE] be achieved?

10. Do you feel we missed something about a European Policy roadmap?

11. In which field do you work?

[42] This question is repeated for each of three objectives answered to 5

12. What is your field of expertise?

13. Where are you from?

APPENDIX 2: CODES FOR THE EXTERNALITIES AND RESEARCH AND INNOVATION TOPICS CONSIDERED

EXTERNALITIES

The following table summarises all externalities considered in the BYTE project. For more details, please see (Donovan, et al. 2014) and (Vega-Gorgojo, Donovan, et al. 2015).

Table 12. Societal externalities considered by the BYTE project.

Code	±	Stakeholders	Main topic	Description
E-PC-BM-1	+	Public sector-citizens	Business models	Tracking environmental challenges
E-PC-BM-2	+	Public sector-citizens	Business models	Better services, e.g. health care and education, through data sharing and analysis (need to explain the benefits to the public)
E-PC-BM-3	+	Public sector-citizens	Business models	More targeted services for citizens (through profiling populations)
E-PC-BM-4	+	Public sector-citizens	Business models	Cost-effectiveness of services
E-PC-BM-5	-	Public sector-citizens	Business models	Need of skills and resources
E-PC-DAT-1	+	Public sector-citizens	Data sources and open data	Foster innovation, e.g. new applications, from government data (data reuse)
E-PC-LEG-1	+	Public sector-citizens	Policies and legal issues	Transparency and accountability of the public sector
E-PC-LEG-2	-	Public sector-citizens	Policies and legal issues	Compromise to government security and privacy (due to data sharing practices)
E-PC-LEG-3	-	Public sector-citizens	Policies and legal issues	Private data misuse, especially sharing with third parties without consent
E-PC-LEG-4	-	Public sector-citizens	Policies and legal issues	Threats to data protection and personal privacy
E-PC-LEG-5	-	Public sector-citizens	Policies and legal issues	Threats to intellectual property rights (including scholars' rights and contributions)
E-PC-ETH-1	+	Public sector-citizens	Social and ethical issues	Increased citizen participation
E-PC-ETH-2	+	Public sector-citizens	Social and ethical issues	Crime prevention and detection, including fraud (surveillance using big data)
E-PC-ETH-3	-	Public sector-citizens	Social and ethical issues	Distrust of government data-based activities
E-PC-ETH-4	-	Public sector-citizens	Social and ethical issues	Unnecessary surveillance
E-PC-ETH-5	-	Public sector-citizens	Social and ethical issues	Public reluctance to provide information (especially personal data)
E-PC-TEC-1	+	Public sector-citizens	Technologies and infrastructures	Gather public insight by identifying social trends and statistics, e.g. epidemics or employment rates (see social computing)
E-PC-TEC-2	+	Public sector-citizens	Technologies and infrastructures	Accelerate scientific progress (improved efficiency in data access, improved data analysis)
E-OC-BM-1	+	Private sector-citizens	Business models	Rapid commercialization of new goods and services

E-OC-BM-2	+	Private sector-citizens	Business models	Making society energy efficient
E-OC-BM-3	+	Private sector-citizens	Business models	Data-driven employment offerings
E-OC-BM-4	+	Private sector-citizens	Business models	Marketing improvement by using targeted advertisements and personalized recommendations
E-OC-BM-5	-	Private sector-citizens	Business models	Employment losses for certain job categories (white-collar jobs being replaced by big data analytics)
E-OC-BM-6	-	Private sector-citizens	Business models	Risk of informational rent-seeking
E-OC-BM-7	-	Private sector-citizens	Business models	Reduced market competition (creation of a few dominant market players)
E-OC-BM-8	-	Private sector-citizens	Business models	Privatization of essential utilities (e.g. Internet access)
E-OC-DAT-1	+	Private sector-citizens	Data sources and open data	Enhancements in data-driven R&D
E-OC-DAT-2	+	Private sector-citizens	Data sources and open data	Fostering innovation from opening data
E-OC-DAT-3	+	Private sector-citizens	Data sources and open data	Time-saving in transactions if personal data were already held
E-OC-DAT-4	-	Private sector-citizens	Data sources and open data	Creation of data-based monopolies (platforms and services)
E-OC-LEG-1	+	Private sector-citizens	Policies and legal issues	Increased insight of goods (more transparency)
E-OC-LEG-2	+	Private sector-citizens	Policies and legal issues	Increased transparency in commercial decision making
E-OC-LEG-3	-	Private sector-citizens	Policies and legal issues	Private data accumulation and ownership (losing control of their personal data)
E-OC-LEG-4	-	Private sector-citizens	Policies and legal issues	Threats to intellectual property rights
E-OC-ETH-1	+	Private sector-citizens	Social and ethical issues	Safe and environment-friendly operations
E-OC-ETH-2	+	Private sector-citizens	Social and ethical issues	Increase awareness about privacy violations and ethical issues of big data
E-OC-ETH-3	-	Private sector-citizens	Social and ethical issues	Invasive use of information
E-OC-ETH-4	-	Private sector-citizens	Social and ethical issues	Discriminatory practices and targeted advertising (as a result of profiling and tracking private data)
E-OC-ETH-5	-	Private sector-citizens	Social and ethical issues	Distrust of commercial data-based activities (due to lack of transparency or unintended secondary uses of data)
E-OC-ETH-6	-	Private sector-citizens	Social and ethical issues	Unethical exploitation of data, e.g. some types of tracking and profiling, encompassing concerns about discrimination and dignity (especially relevant in sensitive domains such as health or finance)
E-OC-ETH-7	-	Private sector-citizens	Social and ethical issues	Consumer manipulation

E-OC-ETH-8	-	Private sector-citizens	Social and ethical issues	Private data leakage (concern about data protection and cyber threats, especially bankcard fraud and identity theft)
E-OC-ETH-9	-	Private sector-citizens	Social and ethical issues	Private data misuse, especially sharing with third parties without consent
E-OC-ETH-10	-	Private sector-citizens	Social and ethical issues	Privacy threats even with anonymised data (easy to de-anonymise) and with data mining
E-OC-ETH-11	-	Private sector-citizens	Social and ethical issues	Public reluctance to provide information (especially personal data)
E-OC-ETH-12	-	Private sector-citizens	Social and ethical issues	"Sabotaged" data practices
E-OC-ETH-13	-	Private sector-citizens	Social and ethical issues	Lack of context or incomplete data can result in incorrect interpretations
E-OC-TEC-1	+	Private sector-citizens	Technologies and infrastructures	Free use of services, e.g. email, social media, search engines
E-OC-TEC-2	+	Private sector-citizens	Technologies and infrastructures	Optimization of utilities through data analytics
E-CC-ETH-1	-	Citizens-citizens	Social and ethical issues	Continuous and invisible surveillance
E-CC-TEC-1	+	Citizens-citizens	Technologies and infrastructures	Support communities
E-OO-BM-1	+	Private sector-private sector	Business models	Opportunities for economic growth through community building (sharing information and insights across sectors)
E-OO-BM-2	+	Private sector-private sector	Business models	Innovative business models through community building (sharing information and insights across sectors)
E-OO-BM-3	-	Private sector-private sector	Business models	Challenge of traditional non-digital services, e.g. new data-driven taxi and lodgement services
E-OO-BM-4	-	Private sector-private sector	Business models	Monopoly creation through the purchase of data-based companies
E-OO-BM-5	-	Private sector-private sector	Business models	Competitive disadvantage of newer businesses and SMEs (creation of a few dominant market players)
E-OO-BM-6	-	Private sector-private sector	Business models	Reduced growth and profit among all business, particularly SMEs (creation of a few dominant market players)
E-OO-BM-7	-	Private sector-private sector	Business models	Increased demand in computing power, data storage or network capabilities
E-OO-DAT-1	-	Private sector-private sector	Data sources and open data	Inequalities to data access (digital divide between big data players and the rest)
E-OO-DAT-2	-	Private sector-private sector	Data sources and open data	Dependency on external data sources, platforms and services (due to dominant position of big players)
E-OO-DAT-3	-	Private sector-private sector	Data sources and open data	Threats to commercially valuable information
E-OO-DAT-4	-	Private sector-private sector	Data sources and open data	Distrust on data coming from uncontrolled sources

E-OO-ETH-1	-	Private sector-private sector	Social and ethical issues	Market manipulation
E-OO-TEC-1	-	Private sector-private sector	Technologies and infrastructures	Barriers to market entry (due to dominant position of big players, the need for major investment, and the complexity of big data processing)
E-PO-BM-1	+	Public sector-private sector	Business models	Opportunities for economic growth (new products and services based on open access to big data)
E-PO-BM-2	+	Public sector-private sector	Business models	Innovative business models (closer linkages between research and innovation)
E-PO-DAT-1	-	Public sector-private sector	Data sources and open data	Open data puts the private sector at a competitive advantage (they don't have to open their data and have access to public data)
E-PO-DAT-2	-	Public sector-private sector	Data sources and open data	Inequalities to data access, especially in research (those with less resources won't be granted access to data)
E-PO-LEG-1	-	Public sector-private sector	Policies and legal issues	Lack of norms for data storage and processing
E-PO-LEG-2	-	Public sector-private sector	Policies and legal issues	Reduced innovation due to restrictive legislation
E-PO-ETH-1	-	Public sector-private sector	Social and ethical issues	Taxation leakages (intermediation platforms, delocalization of data-based corporations)
E-PP-LEG-1	-	Public sector-public sector	Policies and legal issues	EarthObvspolitical tensions due to surveillance out of the boundaries of states
E-PP-LEG-2	-	Public sector-public sector	Policies and legal issues	Need to reconcile different laws and agreements, e.g. "right to be forgotten"

GROUPS OF EXTERNALITIES

The following classification of externalities follows the analysis in (Lammerant, De Hert and Lasierra Beamonte, et al. 2015).

Table 13. Classification of externalities.

	Externality group	Externalities included
Economic externalities	Improved efficiency	E-PC-BM-2, E-PC-BM-3, E-PC-BM-4, E-PC-TEC-1, E-OC-ETH-12, E-OC-ETH-13, E-OC-TEC-2, E-OO-BM-7
	Innovation	E-PC-BM-2, E-PC-DAT-1, E-PC-TEC-2, E-OC-DAT-1, E-OC-DAT-2, E-PO-BM-1, E-PO-BM-2, E-PO-LEG-2, E-OO-BM-1, E-OO-BM-2
	Changing business models	E-OO-BM-1, E-OO-BM-2, E-OO-BM-3, E-OO-BM-5, E-OO-DAT-1, E-PO-BM-1, E-PO-BM-2, E-PO-DAT-1, E-OC-BM-5, E-OC-BM-7, E-OC-BM-8
	Employment	E-OC-BM-3, E-OC-BM-5, E-PO-BM-1
	Role of public funding	E-PO-DAT-1, E-OO-BM-2

Social and ethical externalities	Beneficial impacts due to improved efficiency and innovation	E-PC-BM-1, E-PC-BM-2, E-PC-BM-3, E-PC-ETH-2, E-OC-ETH-1
	Improved awareness and improved decision-making	E-PC-TEC-1, E-PC-BM-1, E-PC-BM-2, E-PC-BM-3, E-OC-ETH-1, E-PC-ETH-2, E-PC-LEG-1
	Participation	E-PC-ETH-1, E-CC-TEC-1
	Equality	E-OC-ETH-4, E-OO-DAT-4
	Discrimination	E-OC-ETH-2, E-OC-ETH-4, E-OC-ETH-9
	Trust	E-PC-ETH-1, E-PC-ETH-5, E-OC-ETH-2, E-OC-ETH-7, E-OC-ETH-11, E-OC-ETH-12, E-OC-ETH-13, E-OO-ETH-1, E-OO-DAT-4, E-PC-LEG-3, E-CC-ETH-1
Legal externalities	Data protection and privacy	E-PC-LEG-3, E-PC-LEG-4, E-OC-LEG-3, E-OC-ETH-2, E-OC-ETH-3, E-OC-ETH-4, E-OC-ETH-7, E-OC-ETH-9, E-OC-ETH-10, E-CC-ETH-1, E-PO-LEG-2, E-OO-TEC-1
	Intellectual property rights	E-PC-LEG-5, E-OC-LEG-3, E-OC-LEG-4, E-PO-LEG-1, E-PO-LEG-2, E-PP-LEG-2
	Liability and accountability	E-PO-LEG-1
Political externalities	Relations between private vs. public and non-profit sector	E-OO-DAT-2, E-OO-BM-5, E-OC-BM-8
	Losing control to actors abroad	E-OC-BM-8, E-PP-LEG-1, E-PP-LEG-2, E-PO-LEG-1, E-OC-LEG-1, E-OC-LEG-2
	Improved decision-making and participation	E-PC-LEG-1, E-CC-TEC-1, E-PC-ETH-1
	Political abuse and surveillance	E-PP-LEG-1, E-OC-ETH-9

RESEARCH AND INNOVATION TOPICS

The following research and innovation topics come mostly from the Big Data Value Strategic and Innovation Agenda (Big Data Value Association 2016), with small modifications by the BYTE analysis.

Table 14. Research and innovation topics.

Research topic	Research code
Data management	
Handling unstructured and semi-structured data	R-DM-01
Semantic interoperability	R-DM-02
Measuring and assuring data quality	R-DM-03
Data lifecycle	R-DM-04
Data provenance, control and IPR	R-DM-05
Data-as-a-service model and paradigm	R-DM-06
Data processing	
Architectures for data-at-rest and data-in-motion	R-DPROC-01
Techniques and tools for processing real-time heterogeneous data	R-DPROC-02
Scalable algorithms and techniques for real-time analytics	R-DPROC-03

Decentralised architectures	R-DPROC-04
Efficient mechanisms for storage and processing	R-DPROC-05
Data analytics	
Improved models and simulations	R-DA-01
Semantic analysis	R-DA-02
Event and pattern discovery	R-DA-03
Multimedia (unstructured) data mining	R-DA-04
Machine learning techniques, deep learning for BI, predictive and prescriptive analytics	R-DA-05
Context-aware analytics	R-DA-06
Data protection	
Complete data protection framework	R-DPROT-01
Data minimization	R-DPROT-02
Privacy-preserving mining algorithms	R-DPROT-03
Robust anonymisation algorithms	R-DPROT-04
Protection against reversibility	R-DPROT-05
Pattern hiding mechanism	R-DPROT-06
Secure multiparty mining mechanism	R-DPROT-07
Data visualisation	
End user visualisation and analytics	R-DV-01
Dynamic clustering of information	R-DV-02
New visualisation for geospatial data	R-DV-03
Interrelated data and semantics relationships	R-DV-04
Qualitative analysis at a high semantic level	R-DV-05
Real-time and collaborative 3-D visualisation	R-DV-06
Time dimension of big data	R-DV-07
Real-time adaptable and interactive visualisation	R-DV-08
Non-technical priorities	
Data-intensive engineers	R-SK-01
Data scientist	R-SK-02
Data-intensive business experts	R-SK-03
Technology standardisation	R-ST-01
Data standardisation	R-ST-02
Establish and increase trust	R-SO-01
Privacy-by-design, security-by-design, anti-discrimination-by-design	R-SO-02
Ethical issues	R-SO-03
Develop new business models	R-SO-04
Citizen research	R-SO-05
Discrimination discovery and prevention	R-SO-06

APPENDIX 3: MAPPINGS OF EXTERNALITIES, RESEARCH AND INNOVATION TOPICS AND SECTORS

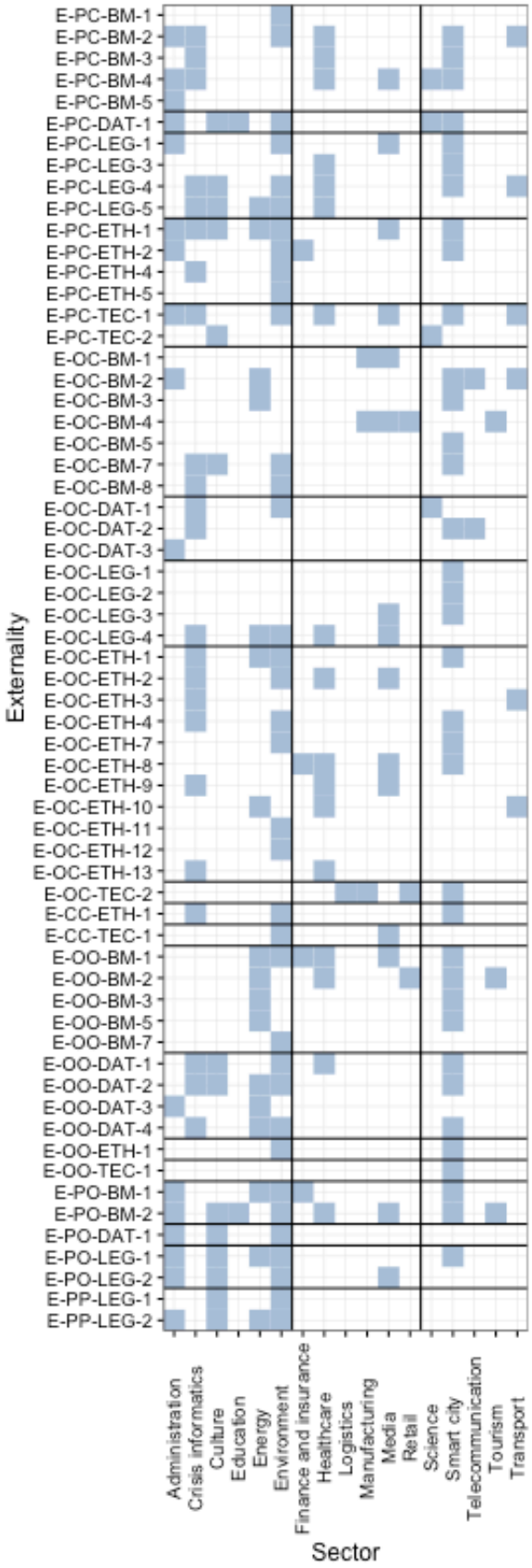


Figure 27. Externalities observed in different sectors, derived from BYTE analysis and the literature review.

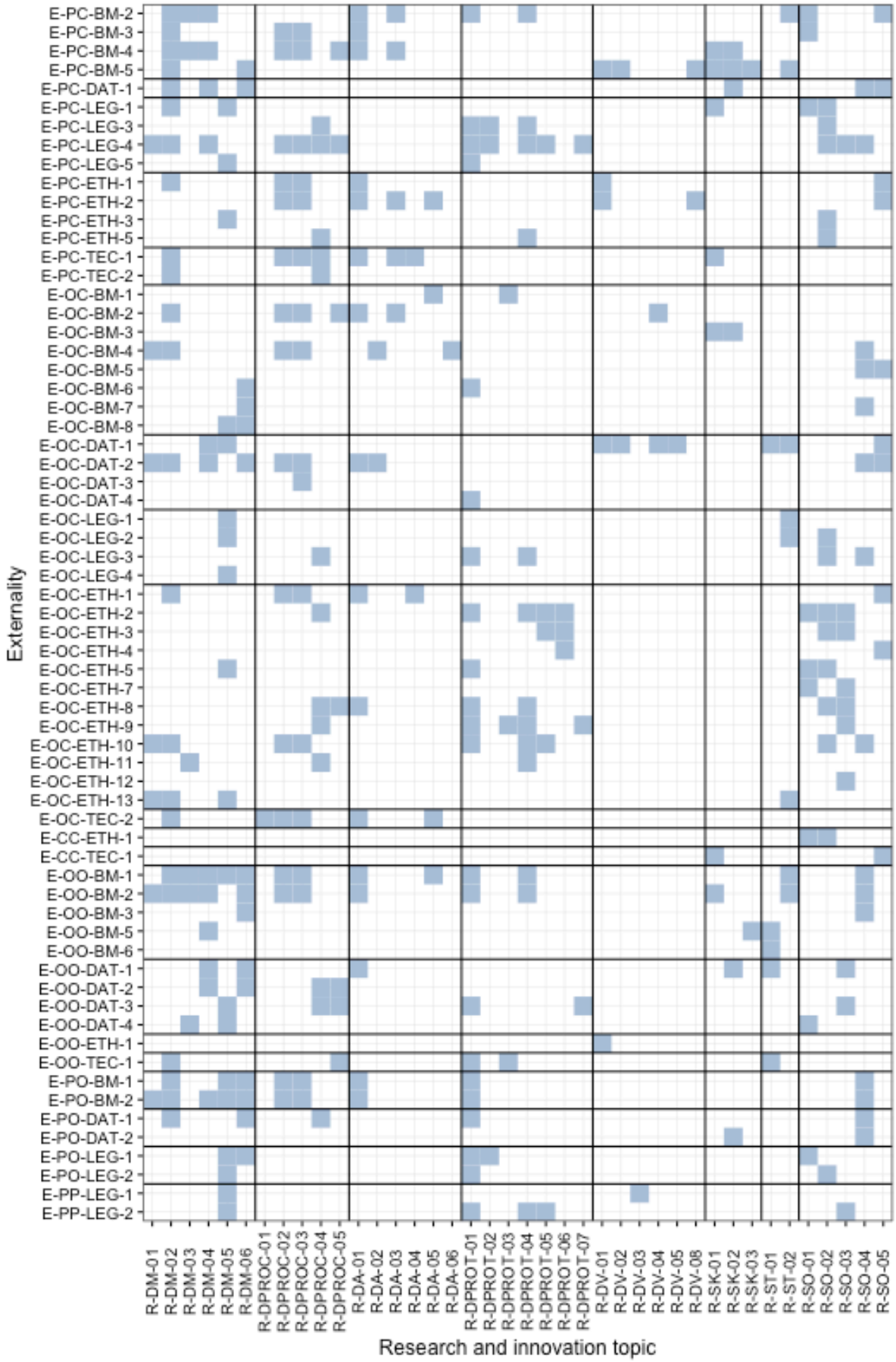


Figure 28. Research linked to externalities, derived from BYTE analysis and the literature review.

APPENDIX 4: PROGRAMME OF THE BYTE BIG DATA RESEARCH ROADMAPPING WORKSHOP

The *Big data research roadmapping workshop* aimed to present, discuss and obtain additional input for the research roadmap from invited participants, in the format of round tables.

The workshop took place on the 1st of July at the Evoluon in Eindhoven, the Netherlands, as a collocated event of the European Data Forum 2016 (29-30 June 2016). It began with a joint session with a parallel workshop on *Big data platforms and benchmarking*, were the three projects BigDataEurope, HOBBIT and BYTE will be presented.

The rest of the workshop was devoted to the research roadmapping exercise. The BYTE research roadmap was presented and participants worked in small groups around round tables, first to discuss and validate research and innovation topics and then to align these topics with the identified societal impacts and externalities. In the second part of the workshop, participants worked on the time-alignment and prioritisation of the research and innovation topics. Finally, the workshop ended with the launch of the BYTE Big Data Community.

Table 15. BYTE Big data research roadmapping workshop agenda.

Time	Topic
9:00 – 10:30	Joint session with Big Data Europe and HOBBIT projects. Three projects in three nutshells.
10:30 – 11:00	Coffee Break
11:00 – 11:15	The BYTE research roadmap. Presentation and exercise description.
11:15 – 11:50	Working groups: Discussion and validation of research topics
11:50 – 12:30	Working groups: Alignment of research topics and externalities
12:30 – 13:30	Lunch
13:30 – 14:15	Working groups: Time alignment and prioritisation
14:15 – 14:45	BYTE Big Data Community launch
14:45 – 15:00	Wrap up

A total of 26 people participated in the workshop of different professional positions, organisations types, industry sectors and countries. Table 16 to Table 19 summarise the participants' profiles.

Table 16. Participants of the research roadmapping workshop by professional position.

Position	Number
Professor	7
Senior researcher	7
Manager	2
Senior professional/consultant	3
Consultant	5

Data analyst	1
Policy and research officer	1

Table 17. Participants of the research roadmapping workshop by organisation type.

Organisation type	Number
Academia	11
SME	8
Large company	3
Public organisation	3
Certification body	1

Table 18. Participants of the research roadmapping workshop by industry sector. There are more sectors represented than participants, due to some participants' activity in multiple sectors.

Sector	Number
Administration/public sector	3
Business intelligence	1
Crisis informatics	2
Culture	2
Energy	3
Environment	1
Healthcare	2
Human computation	1
Law	1
Machine learning	1
Natural language processing	1
Open data	1
Semantic technologies	4
Smart city	2
Social sciences	1

Table 19. Participants of the research roadmapping workshop by country.

Country	Number
Austria	3

Belgium	3
Germany	3
Hungary	1
Ireland	2
Italy	1
Norway	1
Spain	1
Sweden	1
The Netherlands	7
United Kingdom	3