European Commission

BIG DATA VALUE
PUBLIC-PRIVATE PARTNERSHIP

Policy4Data Policy Brief

Big data in Europe for 2020 and beyond:

Policy insights and recommendations from current H2020 big data projects.

Disclaimer

The CDB Policy4Data Policy Brief was co-authored by representatives funded under the European Commission's Horizon 2020 Research and Innovation programme under the following Grant Agreements: BDVe #732630, MyHealthMyData #732907, BigDataStack #779747, EW-Shopp #732590, E-sides #731873, Transforming Transport #731932, BigPolicyCanvas #769623, Lemo-H2020 #770038. The information, views and recommendations set out in this publication are those of the CDB Project Group and cannot be considered to reflect the views of the European Commission.

Table of Contents

Executive Summary

This policy brief reflects current developments within the several Big Data research projects funded under H2020 and, combined with insights from the BDV PPP summit in Riga [1], aims to contribute to ongoing challenges in Europe around the regulation of big data. This policy brief is a product of the Common Dissemination Booster, funded under H2020. The policy recommendations are based on projects participating in the CDB services.

One of the main challenges identified in this policy brief is that of regulating big data. The contributors represent a multidisciplinary set of scholars, researchers and practitioners involved in either implementing big data solutions, researching data policy and governance, or finding technological solutions for implementing data policies. These activities are distinctly different, yet are intrinsically connected through questions of how Europe can maximize big data benefits while simultaneously protecting rights of individuals and companies.

In the policy brief, we draw from a set of insights and lessons learnt that are based on recent H2020 projects around big data development and implementation in different sectors, ranging from traffic and transport to online retail to the public sector and more. The main solutions offered from the projects are (among others) a data governance taxonomy, tools for automated compliance, a Data Asset Marketplace and a roadmap for using big data for policymaking.

The main recommendations are to support integration and interoperability of public administration databases; to support development of data markets and provide guidance on their effective use, to support work on the adoption of privacy-preserving technologies for big data and AI and to promote data-driven policymaking and regulatory automation.

1 https://www.big-data-value.eu/ppp-summit-2019/

1. Topic Overview

1. 1 Topic Overview

Big data and AI are currently top-of-mind themes in many technical-and non-technical debates. Where on the one hand, big data technologies come with a set of large claims and promises concerning its disruptive potential in many, if not all sectors, it also comes with large risks, be they societal, economic or scientific.

Data governance is becoming increasingly important on both a strategic-and operational level for companies, governments and organisations alike, due to the role of data taking more centre-stage in many day-to-day processes and decisions. Going truly "data-driven" is a slow process, often regarded as highly risky, and much innovation in data governance models is emerging to address this issue. The DigiTranScope project is developing a strong taxonomy for data governance.

Sector-specific or cross-sectorial interactions between the availability and the need for data are matched via data marketplaces and/or via particular data resources or platforms that offer datasets or algorithms or specific software for analysing data. Access to data and data marketplaces is crucial for stimulating data-driven (economic) activity, as is confirmed by the EWSHOPP project, Big Data Stack and the TransformingTransport pilot projects.

Governments are also realising the enormous potential of the data available in their respective databases and other sources of data, and are striving to unlock the public value of this data by making their data sources interoperable across borders and administrations.

Such actions are being supported by important European-level initiatives such as the Once Only Principle and the European Interoperability Framework. Using big data for policymaking is a developing area of interest, that comes with promises and challenges: a roadmap for research has recently been developed by the Big Policy Canvas project.

At the same time, landmark European legislation such as the General Data Protection Regulation (GDPR) has made it necessary to ensure that powerful technologies such as big data and AI are implemented under full respect of privacy preservation considerations. Where the MYHD project is offering technological solutions to ensure safe data sharing and patient data control mechanisms via blockchain-based mechanisms, the E-SIDES project has developed a highly insightful gap analysis on why and how technological solutions (PPTs)[2] lack uptake and what we can do about it, while the LeMo project has shown the complexity of the data regulatory landscape, and the connection between different data regulations.

2 Privacy Preserving Technologies

1. 2 Policy challenges

Developing policy for big data and AI, that is, developing strategies and approaches to maximize societal, scientific and economic benefit, is every bit as pressing and challenging as developing and adopting big data and AI technologies themselves. A critical challenge for policymakers is to recognize the crucial importance of data as the "fifth freedom" in the European Single Market and to develop a coherent, consistent concept of the nature of data so that policy can support its effective governance and promote the development of innovative governance models.

Although data marketplaces can thrive on their own, policy formulation can provide much needed support through, for example, the facilitation of cross-border flows and creating transparent, simplified regulation of data rights.

The arrival of the GDPR has unfortunately created a false dichotomy in the minds of too many entrepreneurs and businesspersons who believe that privacy preservation and innovation are incompatible. It is an urgent policy challenge at the highest levels to ensure that this false dichotomy does not take root and slow down the pace of European innovation.

Policy development strives to keep pace with the rapid advance of big data and AI technology, but the complex web of factors ranging from privacy mandates (e.g. user consent) to regulatory frameworks inevitably slows it down. And yet, the technology itself contains the seeds of policy innovation, through the largely unexplored potential of data-driven policymaking, whereby the data itself enables rapid and transparent implementation and monitoring, and AI-assisted policy compliance monitoring.

An informative series of appendices in this Policy Brief provides in-depth introduction and treatment of each of the topics covered in this section.

†

2 Recommendations

2.1 Develop and implement different data governance models

Data-driven digital services cover many areas and sectors and involve a large number of stakeholders along the value chain. Yet successful data platforms seem to develop in a converging manner[3]. Ensure that data silos and economic power due to such silos can be better understood and managed, due to network effects. More research is needed on how we can consider and take on board the multiplicity of stakeholders and how, via for instance the taxonomy on different data governance models, we can better understand the role of data governance in balancing

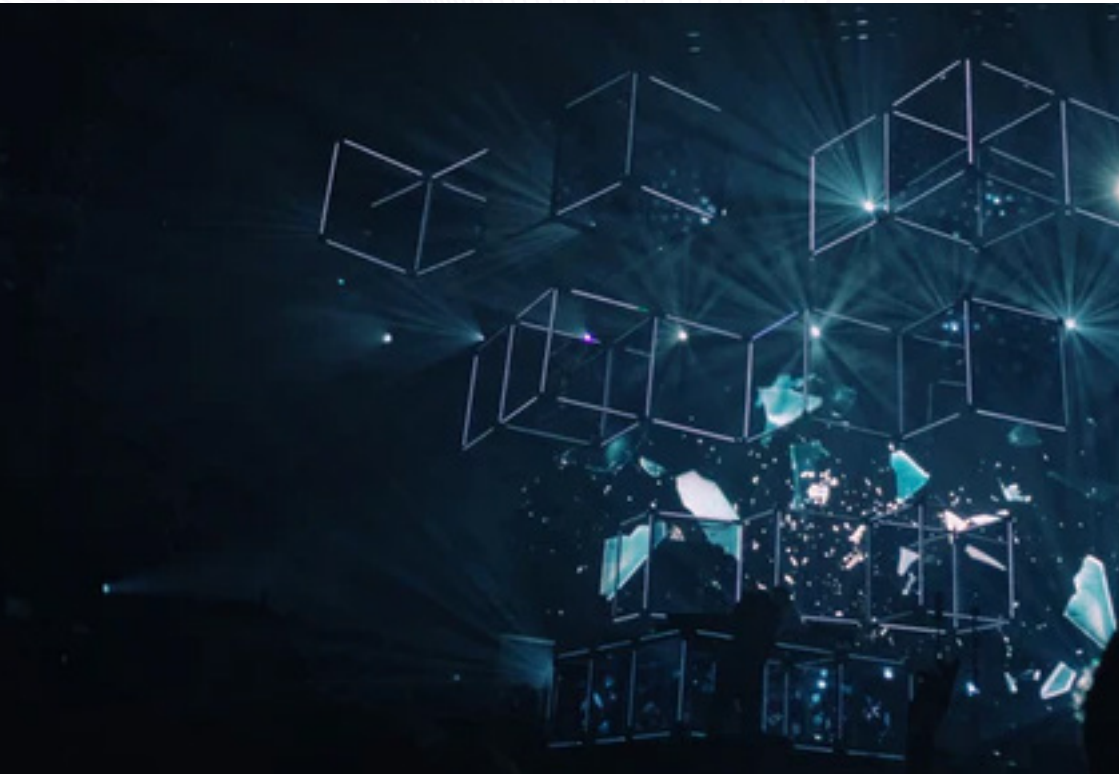
2.2 Support integration and interoperability of public administration databases

The integration and interoperability of government data is becoming increasingly urgent as government holds massive and rapidly growing amounts of data that are dramatically underexploited. In this regard, new solutions are needed that balance the need for data integration with the safeguards on data protection, the demand for data centralisation with the need to respect each administration's autonomy, and the requirement for ex ante homogenization with more pragmatic, on-demand approaches based on the "data lake" paradigm. Data integration has long been a priority for public administrations but with the new European Interoperability Framework and the objective of the once only principle it has become an unavoidable priority. As an example, the Data & Analytics Framework (DAF) by the Italian Digital Team aims to develop and simplify the interoperability of public data between PAs, standardize and promote the dissemination of open data, optimize data analysis processes and generate knowledge to be reused.

3 see http://www.bdva.eu/sites/default/files/BDVA%20DataSharingSpace%20PositionPaper_April2019_V1.pdf

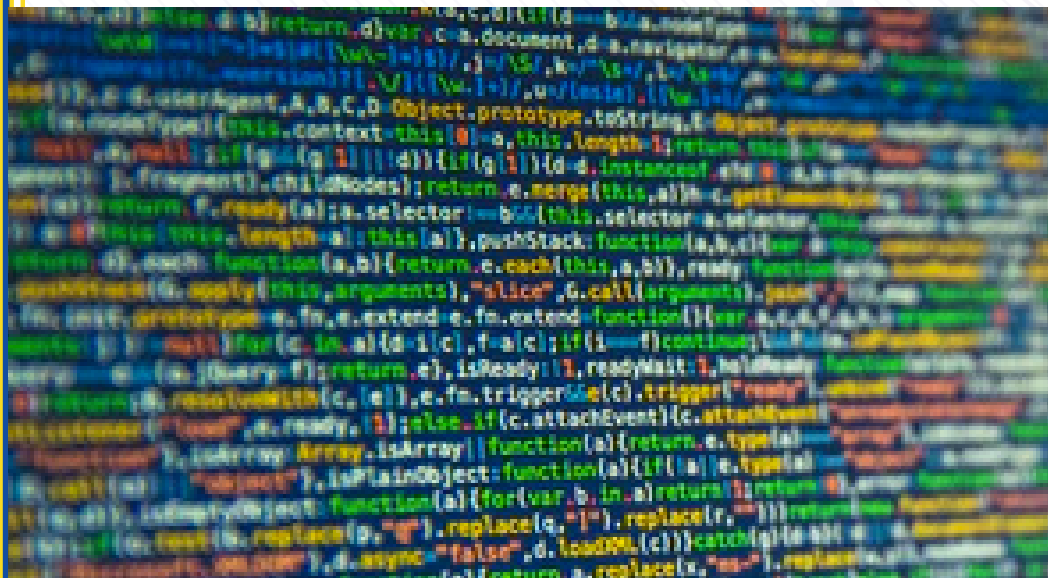
2.3 Support development of data markets and provide guidance on their effective use

Provide policy support for the creation of open data initiatives from different governments in the world. The availability of more data is crucial for organizations and citizens, empowering them to analyse and use these data for a plethora of applications. Support for development of tools for big data management and exploitation is another action which has a great effect in reducing this gap. The first thing to note (even though it might seem obvious, but many times is not well understood) is how requirements vary for different types of applications, stakeholders and organizations. The criticality of the requirements varies greatly. Another important action is to create an awareness of what AI and Big Data are, in what problems and circumstances they can help, and even more importantly, in which cases they are not useful.



2.4 Support work on the adoption of privacy-preserving technologies for big data and AI

The flexible interpretation of privacy and privacy-preserving technologies, which is both a blessing and a curse for professionals designing and using these technologies, could be addressed by policies that offer guidelines on how to insert legal definitions of privacy into design requirements that are tailored to different big data contexts. Policies aimed at bridging differences in EU and US approaches to privacy and competition law could help deconstruct implementation barriers for privacy-preserving technologies. Although US companies handling data of EU residents must comply with GDPR and align US and EU approaches to data protection, the US approach remains quite different. Sector specific policies and best practices for the handling of sensitive data are also perceived as assets by a wide spectrum of professionals. Promotion of collecting and disseminating best-practices would be very helpful.



2.5 Promote data-driven policymaking and regulatory automation

Technology is constantly trying to catch up and provide solutions for organizational changes, which is natural. However, if we would like to make maximum usage of the technology, it would be very beneficial if policies, regulatory frameworks, legislation etc. are written in a machine-readable form that would enable the rapid implementation and monitoring of them. By formulating and describing policies in a way that could be easily transformed into contractual terms, e.g. in smart contracts, we could increase the transparency and the common understanding of the policies from users who are not accustomed to the details of the technology. We need deterministic methods that will be responsible for modelling and storing data privacy policies and user consent. This will be the engine for determining whether data is allowed to be stored, accessed, or transferred based on the owner of the data and the purpose for which it will be used, together with the relevant privacy policies.

3 Project Group

The Big Data Value Association (BDVA) is an industry-driven international not–for-profit organisation with 200 members all over Europe and a well-balanced composition of large, small, and medium-sized industries as well as research and user organisations. BDVA is the private counterpart to the EU Commission to implement the Big Data Value PPP program. BDVA and the Big Data Value PPP pursue a common shared vision of positioning Europe as the world leader in the creation of Big Data Value. Under the Common Dissemination Booster the following projects have joined forces to co-author the Policy4Data Policy Brief.

MY HEALTH MY DATA

www.myhealthmydata.eu/

EWSHOPP

www.ew-shopp.eu

e-SIDES

e-sides.eu/e-sides-project

TRANSFORMING TRANSPORT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731932

transformingtransport.eu/

BigDataStack

Holistic stack for big data applications and operations

bigdatastack.eu/

BIG DATA VALUE eCOSYSTEM PROJECT

www.big-data-value.eu/

BPC BigPolicyCanvas

www.bigpolicycanvas.eu/

LeMO

lemo-h2020.eu/

4 Appendices: Contributions from several projects

4.1 Introduction and General Overview

In a recent EU-wide conference on Big Data, dubbed the BDV PPP Summit 2019, questions on Impact empowered by Data-driven Artificial Intelligence were addressed. During the Summit, several current H2020 research projects joined in a discussion on policy-making for big data to address key challenges for policymakers when it comes to big data and data for AI. The panel took stock of the current lessons learnt and the near future policy challenges for big data solutions, and was structured around the following three themes:

1.Big themes and big challenges
2.Data markets: lessons learnt and regulatory challenges
3.Developments in the regulatory landscape

This policy brief reflects with insights from that session, and it aims to contribute to ongoing challenges in Europe around the regulation of big data. The role of policymakers is to somehow shape-and influence the development of big data and AI according to a commonly understood framework of values. However, the connection between regulator and technology developer, or between law and technology in general, is not always accepted or understood, and values are not always shared.
Taking the GDPR as an example, the often-called argument that it would stifle innovation is, although unjustified, still prevalent in ICT – and business communities. Moreover, recent attempts to call upon the responsibilities of the ICT industry when it comes to the negative effects of ICT are often fiercely contested and combatted[4], and the regulatory landscape around big data is complex.

4 https://www.ft.com/content/4ff9f52c-942e-11e9-b7ea-60e35ef678d2

Some argue a stronger presence and influence is needed from the regulator, where others still argue the market and self-regulation will solve issues such as DeepFakes[5] and winner-take-all monopolies in data markets.

Leaving aside the challenges for the moment of thinking about whether to regulate, another challenge lies in how. A long-standing debate on techno-regulation (meaning: regulation of technology through technology) has recently regained traction due to big data and AI. The automation of compliance, for example, or the use of AI to spot legal harms or to combat fake news, are seen as viable[6], if not the only resort to somehow regulate Big Data and AI. Having Lessig's taxonomy of regulatory powers[7] in mind (being social norms, markets, architecture and law), the question on how to regulate big data and consequently AI, are subject to similar forces and within each of those 4 forces of regulation, there exists a wide variety of challenges; providing a coherent overview of current regulations connected to big data is already a daunting task[8], let alone combining architectural, financial and social dynamics into the mix. Although not claiming to provide all answers to such large challenges, the policy brief aims to put forward some recent developments in this area and to provide

5 See for instance https://www.vice.com/en_us/topic/deepfakes

6 http://www.europarl.europa.eu/RegData/etudes/STUD/2019/624279/EPRS_STU(2019)624279_EN.pdf

7 Lessig, L. (1999). Code is law. The Industry Standard, 18.

8 See the recent Deliverable of the LeMo project for a complete overview: https://lemo-h2020.eu/newsroom/2018/11/1/deliverable-d22-report-on-legal-issues

9 See bdva.eu

Big themes and big challenges

One of the main challenges concerning the regulation of data is its continuous conceptual flux. Where technically we might be able to describe what data is and how it works, socially and culturally, the meaning of data and value we attribute to big data is a moving target. Sometimes coined as the fifth freedom that constitutes the European Single Market[10] (besides goods, capital, services and labour), the idea that data 'wants to be free' and is something that effortlessly moves around and can be used and reused across contexts and services seems not to correspond with big data practices, in which processes of gathering, cleaning and selecting useful data and managing contracts surrounding data are labour-and resource intensive endeavours. But, if data is not free, and/or cannot be seen as similar category compared to goods or services[11], then how can we define it otherwise?

10 See f.i. https://ec.europa.eu/digital-single-market/en/free-flow-non-personal-data
11 See for example Kitchin, R (2014) on the big data revolution.

The need for a definition or at least a better grasp of how data is being 'lived' would help tremendously in developing ways to govern data. The importance of data governance becomes even more apparent when thinking about the need for good data when deploying AI. Data governance can be defined as: "The kind of decisions made over data, who is able to make such decisions and therefore to influence the way data is accessed, controlled, used and benefited from[12].

12 Definition provided by Marina Micheli. JRC

Another challenge for big data and AI is to ensure that enough (and) relevant data is available to train AI and to make useful and relevant data services. Such services need to be both competitive and human-rights based. Whereas many companies and governmental institutions experience data protection rights such as the GDPR blocking innovation, more best practices and ways of doing 'good' data and AI are needed in order to show that this is a false dichotomy (utility versus privacy).

Moreover, data and/or algorithms need to move around and to be reused if we want to maximise the one particularity of data: that it is, or can be a non-rivalrous good. The combination, and matchmaking of data sets and of data with certain algorithms, however, can make data rivalrous and of value. Trading and matchmaking data in data marketplaces in that regard, is another big challenge, as is how to regulate such markets.

4.1 Introduction and General Overview

One of the ways to regulate data is through the balancing of supply and demand: to facilitate, influence and shape data markets. Sector-specific or cross-sectorial interactions between the availability and the need for data are matched via data marketplaces and/or via particular data resources or platforms that offer datasets or algorithms or specific software for analysing data. Access to data and data marketplaces is crucial for stimulating data-driven (economic) activity. Data marketplaces are complicated to form: the one offering the dataset has to make sure the data is tradable, is of sufficient quality and completeness, that the meta-data and dataset description is useful and attractive, that the rights for re-use are transferable, that liability clauses are arranged etc. From the demand-side, similar issues emerge: the monetary risk versus the actual business value of acquiring data needs to be assessed, data needs to be stored and workable, skills and expertise of a particular type of data and/or the context of data is necessary, the data rights for types of reuse need to be clear and policies attached to the data need to be machine-readable etc. For a regulator, there are many moments in the data-market sequence of interactions in which a shaping force or role can be exercised. Think of stimulating cross-border data flows, alleviating access barriers and/or simplifying data - exchange and through access & data ownership rights. There are also risks for the regulator or policymaker: one being that well-intentioned regulation can have an adverse effect, for instance on the protection of personal data, which lead to even larger data monopolies and data silos[13].

13 www.forbes.com/sites/forbestechcouncil/2018/06/26/gdpr-will-make-big-tech-even-bigger/#32047d292592

Data markets in that regard are far from abstract ideas or entities: compared to financial markets, for example the traded good is more actual, although, as stated earlier, it does not necessarily consist of goods (as data is copy-able, reusable, usable in different contexts and for different purposes). Within the BDV PPP, many projects have encountered or are encountering the challenging exercise of valorising their data and thinking about how and when data and/or data models can be considered valuable assets for particular markets or applications. When thinking of markets, a known division is between b2c and b2b but also in c2c, b2g, g2b etc. Within the data landscape, such a division of markets can also be made - yet other forms of classifying are possible as well. Different models exist between verticals (sectors) and horizontals (the data science/ ICT industry) itself[14]. Recent research by the BDVe, in which data-driven startups were analysed, shows that most startups have a b2b approach. For them, getting access to data and managing data and IP- rights are the main challenges. Some startups are in the data matchmaking market themselves[15], whereas others would benefit greatly from better access to -and matchmaking between - databases and their respective owners. Initiatives such as International Data Spaces[16]are of crucial importance here for data access and exchange.

14 Hartmann, P. M., Zaki, M., Feldmann, N., & Neely, A. (2014). Big data for big business? A taxonomy of data-driven business models used by start-up firms. A taxonomy of data-driven business models used by start-up firms.

15 See for instance https://www.dawex.com/

16 See https://www.internationaldataspaces.org/. See also https://ec.europa.eu/futurium/en/system/files/ged/industrial_data_space.pdf

Many pilots raised the issue use of open data being necessary for the offerings of new services or to generate research. Additional further assistance from the EC and the national authorities is required in educating the domain(s) stakeholders on: the understanding of what is open and big data, the value of open and big data how we can monetise its use and develop new business models and to assist them to think more openly on sharing information. One solution currently being developed is a Data Assert Marketplace[17]. Where this seems promising, implementing a Data Asset Marketplace (DAM) requires new architectures, technologies and concepts which will drive a data economy.

17 A data asset is the result of taking the raw material from the run-the-business data and producing higher-quality-data end products to integrate the business and monitor the business. Your data warehouse team should have the mission of providing high-quality data assets for enterprise use. (see https://www.dummies.com/programming/big-data/engineering/data-warehousing-what-is-a-data-asset/Z)

Data regulation landscape

Organizations encounter legal barriers when managing big data. Just as with security and privacy, companies usually do not possess knowledge about the laws related to data protection. To make matters worse, it is also hard to find experts in this topic, because laws change according to the country that the organization works in and laws and regulations are also continuously evolving and changing. GDPR has brought big changes, and it has made a favour to companies because it unifies the laws that European companies must follow, instead of adapting to each different country. However, it is still not well understood and several companies are now afraid of managing data because of the hefty fines that GDPR imposes.

Many contributions in this policy brief refer to challenges in working with data as a result of the GDPR. Some reasons for concern can be found in an uneven uptake and enforcement of the regulation among the different Member States. Although the GDPR does not allow for much leeway when it comes to MS implementation acts, the fact that this happens asynchronously causes (legal) uncertainties.

Moreover, many SMEs in for instance online marketing, are having a hard time dealing with organising data flows between many different parties and with separating personal-from personal data and (often collective) risks to (personal) data rights. Whereas many technological and organisational solutions are currently being developed, on top of already existing ones[18], there are various reasons for the lack of uptake, as analysed in detail by the E-SIDES[19] project.

Among others, there are data protection concerns in Big Data and how much data can be exposed so as to make it impossible to reveal any hidden identities or identification of subjects. At the same time we face the constant problem on how to monitor and enforce in mass scale the compliance with regulation and policies. The heterogeneity of data sources presents hurdles to data consumers to identify the datasets that would generate value and to businesses to incorporate them into their business processes. In this light, there are two trends to mention in regulating data: the first being automation of regulation and policy to ensure compliance, via for instance equipping datasets with sticky policies and link official policies via semantic interoperability; the second being using big data for policy development, via for instance policy-labs and data-driven policymaking initiatives[20].

18 https://e-sides.eu/resources/deliverable-41-results-of-the-gap-analysis
19 idem
20 https://roadmap.bigpolicycanvas.eu/

4.2 Data Governance Models

Presented by Marina Micheli, JRC

Reporting on a recent JRC project (DigiTranScope[21]), Marina Micheli provides definitions of both governance and data governance: "Governance is not government per se, it refers to the different stakeholders involved, their power to make and implement decisions, and to make their voice heard" whereas specifically focussed on data, governance is explained as: "The kind of decisions made over data, who is able to make such decisions and therefore to influence the way data is accessed, controlled, used and benefited from". The concept of data governance is of relevance when discussing policy on data, because different types and modes of governance can lead to very different outcomes for the different data actors in an ecosystem[22].

In the DigiTranScope project, one of the aims is to develop a taxonomy of types of data governance models, based on empirical evidence, and guided by two lines of inquiry: the first being how power relations between stakeholders are established in the different data governance approaches and the second being what value is pursued in different approaches to data and what arrangements are set in place to generate it?

21 https://ec.europa.eu/jrc/communities/en/community/digitranscope

22 See f.i. Zillner, S., Gomez, J. A., Robles, A. G., Curry, E., Södergård, C., Boujemaa, N., ... & Despenic, M.(2018). Data-Driven Artificial Intelligence for European Economic Competitiveness and Societal Progress: BDVA Position Statement, November 2018.

Where questions around data governance often remain vague, the aim of this project is to group similar types by looking at actors and types of actions around data; they distinguish the following:

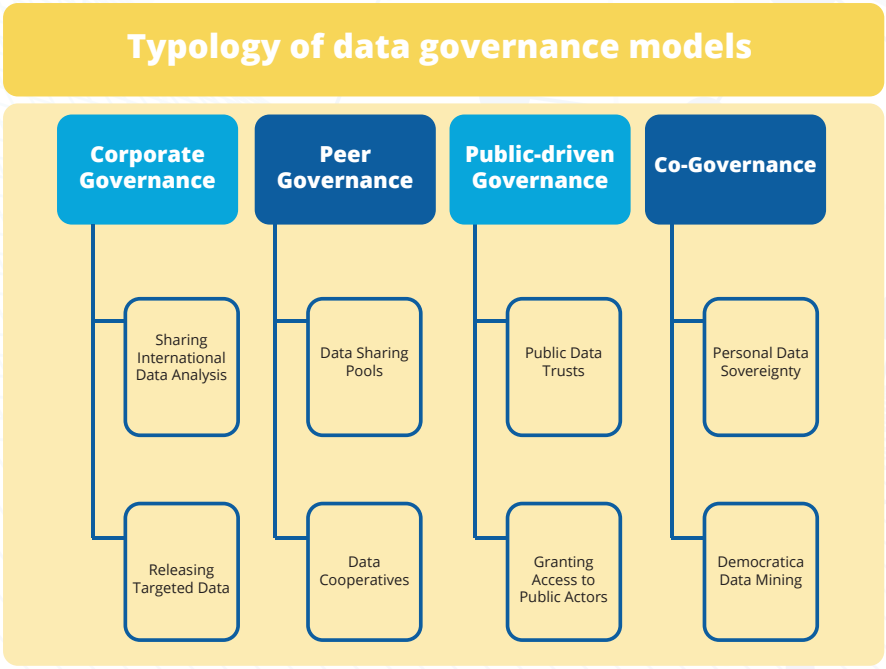


Figure 1: Typology of data governance models

These types, based on desk research and the collection of many examples, are explained in the following table.

Type	Subtype	Description
corporate governance	sharing internal data analysis	**Data platforms,** mobile operators and big players share insights, not data, with external actors
	releasing targeted data	**Data platforms, mobile operators and big players grant access to a restricted number of selected trusted parties (e.g. researchers, NGOs) to specific data subsets that can be used according to platforms' conditions.**
peer governance	data sharing pools	**Data platform companies partner with external organisations to share, combine and analyse each other's data by filling knowledge gaps and reducing duplicate effort.**
	data cooperatives	**Citizens voluntarily pool their data together to create a common pool for mutual benefit. *This model attempts to rebalance the asymmetric relation between data subjects and data users by re-distributing value created from data.**
public-driven governance	public data trusts	**Public bodies, especially local governments, assume a leading role for the aggregation and analysis of citizen data. *Data is understood as a key infrastructure of a city. Private data of public interest is demanded to be accessible for public good.**
	granting access to public actors	**Facilitate access of public bodies to privately-held data of public interest**
co-governance	personal data sovereignty	**Users have greater control concerning which personal information to share, with whom and for what purposes. *Data subjects have more power to decide about their own information compared to what normally happens today on digital platforms.**
	democratic data mining	**Redistribution of power over data analysis among users.**

Table 1: Characteristics of data governance models

Such a taxonomy, although in need of more empirical testing, provides a firm basis to (re) think and redress power asymmetries between big data platforms and other actors. Data-driven digital services cover many areas and sectors and involve a large number of stakeholders along the value chain. Yet successful data platforms seem to develop in a converging manner[23]. Making sure that data silos and economic power due to such silos can be better understood and managed, due to network effects. More research is needed on how we can consider and take on board the multiplicity of stakeholders and how, via for instance the taxonomy as addressed above, we can better understand the role of data governance in balancing different data interests.

4.3 Data markets created out of consumer data

Presented by Theodora Varvarigou, National Technical University of Athens

Introduction

The exploitation of personal consumer data is key to the development of products and services that add value to the business e.g. purchasing recommendations, personalized medicine and lifestyle management, urban mobility (e.g., traffic behaviour modelling) and more. Data markets created out of consumer data are emerging around the new business trends making data be represented as a commodity nowadays.

23 http://www.bdva.eu/sites/default/files/BDVA%20DataSharingSpace%20PositionPaper_April2019_V1.pdf

Emerging privacy concerns in Big Data, associated with the processing of personal/consumers' data and the lack of adequate incentives to overcome these deeply personal concerns, can prevent the deployment, operation and wider use of such services. When such services are developed without the implementation of proper privacy enforcement, businesses are putting themselves at risk. Both citizens and enterprises have to deal with a balance between innovation, participation and legal/regulatory compliance.

The way ahead for Privacy in Data Marketplaces is based on a set of specific needs. In particular for new toolsets, techniques, best practices and policies, which allow enterprises to formulate and resolve trade-offs associated with the legitimate and profitable use of personal data in the scope of providing innovative services based on consumer data analytics. Moreover, there is a need for a new business context to treat "data as commodity": policies and regulations needed as in stock markets, securities markets etc.

However, there are blocking factors which would hinder such development. There are privacy concerns in Big Data and how much data can be exposed so as to make it impossible to reveal any hidden identities or identification of subjects. At the same time, we face the constant problem on how to monitor and enforce in mass scale the compliance with regulation and policies. The heterogeneity of data sources that presents hurdles to data consumers to identify the datasets that would generate value and to businesses to incorporate them into their business processes.

Main development/big data solution (app, service, platform) and for which sector

The way towards introducing concepts for implementing a Data Asset Marketplace (DAM) requires new architectures, technologies and concepts which will drive a data economy by linking sellers to buyers giving the appropriate value, context and quality to data and their usage ensuring ownership and privacy wherever and as much needed. There is a need for blockchain-supported architectures which will ensure that no data transaction will take place without being recorded and will allow at the same time new means for monetizing the data content. Smart contracts can contribute also in the implementation and monitoring of policies through "programmable regulations".

Privacy preserving techniques are needed to be spread vertically in the overall architectures guaranteeing that the data will be valuable only for their intended purpose. The goal of these tools will be to infuse privacy guarantees in novel big data applications/services by applying state-of-the-art privacy preservation mechanisms, properly adapted to the "big" nature of the data.

Advanced mechanisms for data virtualization are needed, that will allow the exposure and discovery of datasets and their properties to potential data consumers/buyers. These new mechanisms will have to separate the "logic" of finding data from the "logic" of using data, enabling innovative companies to develop data-intensive applications that need to consume data from a variety of heterogeneous information sources via a request to a single access point, regardless of data location.

The Data Assets Marketplace will have to rely on a suite of technological layers and is facing significant technical challenges that have to be handled. Data privacy risk assessment techniques will have to describe the data privacy risks in a quantitative and human intuitive way. These will enable the citizens to better estimate possible dangers and harms they may suffer if they offer their data, before they decide to do it. From the enterprise's perspective the techniques will have to aim at providing a decision support tool for deciding whether the aggregation of data under different contracts (agreed consent terms) and privacy policies can lead to a high risk of privacy loss.

As for the Privacy Policy and Consent Management, we experience that legislation and privacy norms are becoming increasingly strict: Any data access attempt for which there is no consent for the specified purpose should be blocked. For this reason, we need deterministic methods that will be responsible for modelling and storing data privacy policies and user consent. This will be the engine for determining whether data is allowed to be stored, accessed, or transferred based on the owner of the data and the purpose for which it will be used, together with the relevant privacy policies. Blockchain - based decentralized storage will serve the purposes of integrity and traceability of data.

Under smart contracts we will be able to have the necessary actions on "programming" and "supervising" policies and regulatory frameworks. Blockchains, however, have their limitations and cannot be regarded as a global solution for privacy as their content is often publicly available to the participants in the blockchain network.

Therefore, it is a common trend nowadays to use blockchains only to provide a timestamp for data that are held off-chain. If specific content needs to be taken down from a public source, the fact that the content existed at a given point would still remain in the blockchain, but the stored hash would now point in another content (or not content if removed)

Data transformation mechanisms responsible to apply specific techniques for data encryption, as well as delivering the data using privacy enforced methods will have to be more sophisticated and agile. Mixture of techniques such as Zero Knowledge Proof (ZKP), Differential Privacy for complex aggregated data and Data Fuzzification are needed and be able to be used in a human intuitive way.

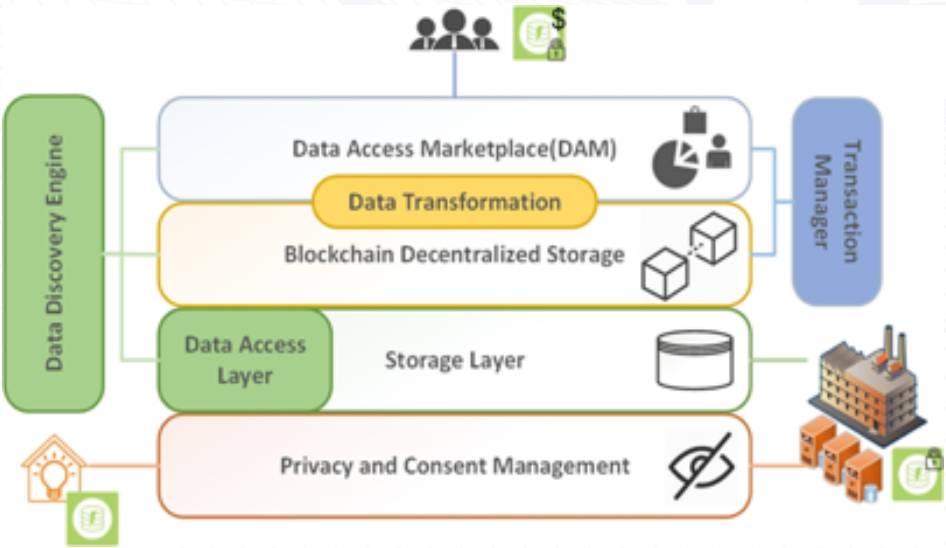


Figure 2: Layered structure of DAM

We see mainly the trend that technology is trying to catch up and provide solutions for organizational changes and this trend is, of course, a natural norm. However, if we would like to make maximum usage of the technology it would be very beneficial if policies, regulatory frameworks, legislation etc. could be "coded" in an appropriate way that would enable the rapid implementation and monitoring of them. By formulating and describing policies in a way that could be easily transformed into contractual terms e.g. in smart contracts, we could increase the transparency and the common understanding of the policies from users who are not accustomed to the details of the technology.

4.4 GDPR compliant health data shareability

Presented by Edwin Morley-Fletcher, Lynkeus

Introduction

MHMD is an H2020 EU-funded research and innovation project which aims at making it possible to either directly share health data, or provide computation outcomes derived from those data, in a fully GDPR-compliant manner. The project is focussed on developing ways to make health data accessible in a safe, secure and privacy-friendly way. Whereas the project is technology-driven, meaning, it aims to develop technological frameworks and solutions to share health-data, the interface- and link with data management and data governance is obvious, as the developed solutions are developed as a response to, and in connection with data regulation and data governance developments in the healthcare sector.

On data sharing: According to various circumstances, and privacy-preserving needs, the health data are shared either as pseudonymous or anonymous data. A semi-automated tool, AMNESIA, is used for providing the necessary pseudonymisation or anonymisation. Alternatively, it is also possible to share data as synthetic data.

The blockchain

Records what transactions happen among the participants to MHMD specifies under what conditions with what permission system with what type of consent the authorised access to the data can be safely enacted through the relevant smart contracts applying the appropriate privacy preserving technologies. The first architectural precondition is a decentralised off-chain storage, by which all raw data is left where it originally belongs, within the repositories of individuals and clinical centres.
Any registered user can browse and analyse what types of data are findable in MHMD through the harmonised description generated by the comprehensive metadata Catalogue. This Catalogue is precisely aimed at ingesting, indexing and discovering all the needed metadata. So as to foster a streamlined, integrated and homogeneous system for dataset search Providing statistical representations and analytics, with appropriate privacy guarantees. Output privacy ensures that sensitive information is not revealed within the Catalogue's queries results.

A scenario:

A researcher browses MHMD Catalogue Finds out that within the system there are datasets corresponding to her research needs, and formalizes her request to have those specific data published or computed. The MHMD blockchain system is the digital space where the execution of the researcher's request can be enacted on a new layer of automation: the process automation based on Smart Contracts.

On synthetic data: Pseudonymised data are re-identifiable and need purpose specific consent. Anonymised data need to be fully non-re-identifiable and therefore risk becoming poor in the information they convey. Synthetic data are fully artificial data, automatically generated by making use of machine learning algorithms, based on recursive conditional parameter aggregation, operating within global statistical models. They retain significant information usefulness. They belong to no really existing persons. By definition, they do not allow any personal re-identification of original individual datasets. They do not fall within the scope of the GDPR. They are freely tradeable.

On differentially-Private Synthetic Data Generation: Adding appropriate differential privacy features can further assure non-reidentification even on whole population statistics. A scalable quality-control system allows to generate synthetic data even more informative and robust than original ones. Quality control and iterative approaches can lead to statistically equivalent sets, at a vastly lower cost.

Such methods can also enrich the synthetic set with more statistical features and, in the case of synthetic images, with automatically placed annotations to then train diagnostic image recognition systems. MHMD has proved that high-quality synthetic cardiovascular images have been generated by using Generative Adversarial Networks (GANs).

Secure Multiparty Computation is a type of cryptography which allows parties to jointly compute a function over their inputs, keeping these inputs private. SMC allows a set of distrustful parties to perform the computation in a distributed manner, while each of them individually remains oblivious to the input data and the intermediate results. The computation is considered secure if, at the end, no party knows anything except its own input and the results.

Homomorphism is the generic property of an encryption scheme which allows performing operations directly on encrypted data. Several existing encryption schemes are available which are homomorphic to any or certain operations. One Partial Homomorphic Encryption solution has been developed within MHMD by the Transylvania University of Brasov, and was showcased by the EU Innovation Radar. This solution is still being tested and presented in scientific publications. It has not been integrated in MHMD, though regarded by all reviewers as a meaningful contribution to the field.

Learning from data:

A deep learning model is employed to work on homomorphically encrypted input-output data. This deep learning model outputs encrypted results which the researcher can decrypt with the symmetric key. The secure distributed processing of the sensitive data is thus performed in such a way that no one party learns anything about the data, nor the other party about the machine learning model. Both data and predictions remain private and data analysis is performed only on the encrypted version of the data. MHMD is using SMPC and Differential Privacy (DP) in the context of a "black-box" federated learning framework, in which a secure ML request containing a model training pipeline is distributed to the data providers along with a set of parameters, and is run locally on an isolated environment. Local computation results (e.g., model gradients) are then securely aggregated using the MHMD SMPC engine (based on the open source Scale-Mamba library). This cycle is repeated to obtain many training iterations and/or model validation.

Summary

Privacy by design is a fundamental characteristic of MHMD, together with process automation and transaction costs minimisation. Researchers can see what types of data are on offer in MHMD without needing to access the data. Data can be published in three formats (pseudonymous, anonymous, synthetic). Secure computation, which permits running AI without disclosing neither data nor algorithms, is performed through SMC and HE. An overall MHMD Privacy-by-Design and GDPR Compliance Assessment is currently being finalised.

4.5 Data integration and enrichment in the marketing industry

Presented by Fernando Perales, JOT Internet Media

Introduction

Digital marketing represents an illustrative example of an industry that has evolved from a creative and design basis to a data-driven approach. The budget limitation to invest in marketing, align with the lack of knowledge in this domain by SMEs, motivate the collection of as much as possible KPIs (such as clicks, impressions, conversion rates, date, time and so on) and the development of automation tools optimising how, where and when the marketing budget is invested.

In this scenario, JOT INTERNET is developing an innovative solution within the EWSHOPP project[24]. The core motivation is the need of achieving higher impact indicators in the marketing campaigns while keeping stable the level of investment. The pilot is also based on small scale examples showing that user interests depends on external factors like weather conditions and relevant events (like calendar events). For that reason, the main goal of this business pilot is the integration of marketing performance indicators with weather and external events data to implement a set of marketing services from the schedule for marketing campaigns launch to the prediction of the impact in ongoing campaigns based on weather forecast and calendar events.

24 https://www.ew-shopp.eu/

The generation of such as services requires the development of tools in all the data value chain:

1.Data collection: Access to JOT internal data, related to all the variables defining the performance of the historical campaigns at keyword level (the lowest possible) and collection of external data, in this case, from EWCMF[25](for weather data) and event registry[26] (for external events)

2. Data preparation: In order to be able to generate the expected insights, all data sets need to be prepared and linked properly. This set of tasks has represented a significant amount of time and effort, approximately 60-70%. In this case the project developed ASIS and Grafterizer tools, enabling the enrichment, transformation, semantic annotation and hosting of the data. It was required dedicated effort on understanding the business motivation and the data to agree on the aggregation level as well as the linking variables allowing the enrichment process. Data preparation has been carried out in collaboration with SINTEF and UNIMIB, as core technical partners of the project for this business pilot.

3. Analytics: This process is supported by a combination of QMINER library and dedicated functions. The goal is the modelling of the keywords based on their correlation with the weather and the external events. In this task the required machine learning algorithms used are word embeddings and clustering, enabling the semantic aggrupation of the keywords following the Google taxonomy and reducing the number of models needed for service implementation.

25 https://www.ecmwf.int/
26 https://eventregistry.org/

4. Output and Visuals: For the results it is needed a combination of data (mainly in a *.csv format) enabling the connection with JOT internal automation tools and visuals supporting the account manager to monitor the results and evaluate the quality of the solutions, as well as prepare the workload for the next period (typically the following week). For the visuals, the business case is supported by the Knowage solution (developed by project partner Engineering)

5. Business exploitation: For JOT, as a business partner, all the technical developments have to be motivated by a business objective. In this case, the generation of new services integrating these external data sources enables the company to boost the impact of the campaigns, investing in those ones formed by the keywords the society are more interested in based on the environmental conditions.

From the above paragraphs it is proved, that, although the value chain is standard, the solutions developed are highly customized to the business case and is particular conditions in terms of data and expected goals.

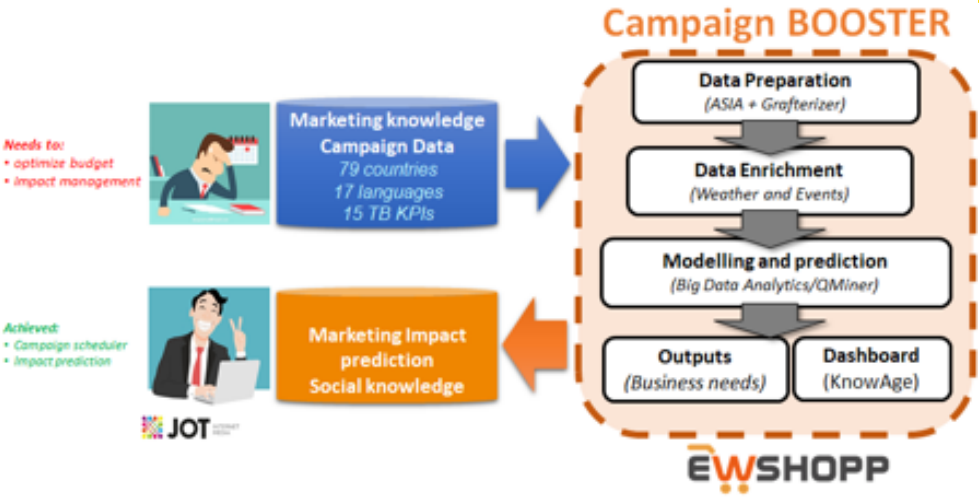


Figure 3: Main components of the business pilot

Main lessons learnt / challenges

This business pilot highlights which are the main technical challenges when dealing with a BigData and machine learning related problem and how it can be solved. From this experience the collaboration among all experts in the data value chain is the only approach to ensure reaching the expected goals. It is not possible at all for a single partner to deal with all the technical and business aspects. For example, JOT has some experience in how to collect the data from the marketing platform and how the data can be stored in the cloud to facilitate the access, sharing and storage. Also, it has a very clear view about the insights needed to implement new services demanded by the market, however, as an SME, JOT has not the resources nor expertise to develop the data enrichment and analytics needed. For that reason, the participation in collaborative project like EWSHOPP is the best way to:

1. Develop the required solution in close collaboration with top partners at European level

2. Increase the company background and know-how related to the data management value chain based on the pilot case and the work developed by the partners.

Therefore, we consider that being open about communicating the business goals, discussing with experts the technology challenges, as well as, sharing the data to work on them and enabling the generation of new knowledge and services is the only way to really implement a data driven culture in the private sector, as JOT does in the EWSHOPP project.

Connection to next project

Taking into consideration the fact that many of the data preparation and enrichment challenges have being already solved, the next project can be focused on the implementation of an Artificial Intelligence toolkit as a Service (AIaaS). This will support the marketing account managers to process the data sets based on their own criteria, as they are the ones with more knowledge about marketing, enabling the definition of innovative strategies to optimize the campaigns and keep on increasing the competitiveness of the company.

4.6 Policy recommendations coming from TT pilots

Presented by Vivian Akrivi Kiousi, Instrasoft

The Traffic & Transport lighthouse project[27] is in its final stages, and in its many pilot studies, lessons can be drawn from the pilot projects.

Concerning the GDPR

Pilots came across fragmented policies regarding GDPR across Europe. My stakeholders were hindered to share data, making big data analysis and use difficult and sometimes not possible. Pilots did follow specific methodologies to facilitate this which delayed their business. Push the EU member states to adopting GDPR at the same level since until now we don't have the same level of adoption; Extra training or the inauguration of assistive tools was suggested by pilots; Natural language explanations to be offered for everyday users as current guidelines are stiff and too legal oriented (i.e. via an online tool).

When it concerns data collaboration, there is an expressed need for the authorities to become more alert on cases where GDPR weakens competition and competitiveness, and in these occasion authorities could direct lawmakers to not hesitate to make necessary adjustments for helping business. Pilots have suggested that national or regional authorities be the ones interpreting complicated issues such as: who owns the data and which data are personal.

27 https://transformingtransport.eu/

More actions should be taken to foster data and more guidance or definition to come from higher level authorities on how data should be stored / used etc. Data Integrity issue (need for regulation to push stakeholders on the type of data they provide across platforms and ensure that these data are reliable and of good quality). Standardisation issues mentioned by pilots: issue of data digitization is mentioned several times for cases where not all data follow the necessary format required by big data technologies. type of data they provide across platforms and ensure that these data are reliable and of good quality). Standardisation issues mentioned by pilots: issue of data digitization is mentioned several times for cases where not all data follow the necessary format required by big data technologies.

TT experts can contribute on working groups to work on the harmonization of the high value datasets to make data economy more efficient - specifically to contribute at W3C and Joinup ISA https://ec.europa.eu/isa2/home_en activities and together with other experts to work on the common structure.

Many pilots raised the issue use of open data being necessary for the offerings of new services or to generate research. Additional further assistance from the EC and the national authorities is required in educating the domain(s) stakeholders on: the understanding of what is open and big data, the value of open and big data how we can monetise its use and develop new business models and to assist them to think more openly on sharing information.

Specific examples

In Airports and railway companies/stakeholders are hindered into opening their data since they consider that such data reveals information to their competitor. Ports expressed different opinions depending on the type of organisations involved and business at stake. On the current document worked by the expert Group TT suggests that governments act as a neutral place where all data sharing happens and since they have the strength through regulation to decide on data handling for appropriate use.

Specific examples

In the TT urban pilots, the need for data sharing has been demonstrated. Companies that won a concession – (public contracts) do not like to share their data with others. If these data become available, via government push, cities can understand better the logistics dynamics and be in the position to analyse traffic flows and do better handling of traffic. In the TT pilot for railway, Thales had to do a special agreement to use weather data that are owned by a company operating in the station. The Data Market Economy should move to a structure where agreements and sharing becomes easy to understand.

Benefits of collaboration emerged identified: Academia provides the theoretical framework of big data use that can be applied and tested (which brings the added value to companies). Academic partners benefit by getting access to real world projects, and the opportunity to evaluate theoretical knowledge in practice through specific applications, when they work with partners coming from industry. Despite the fact that many Universities have established technology transfer offices for the big data matters, they still don't have direct contact with the market to get the required deep understanding of the market needs as well as such a high level of interaction with the stakeholder. The industry does not have the resource to invest in research and their bridge to innovative breakthroughs is the Academic world.

TT demonstrates results to create trust from the industry side to push the big data use via being open to new capabilities, foster the shift of regulation to incorporate big data in several processes. So far things are rather strict and change is not coming fast enough to allow the fast adoption of big data. And some domain specific recommendations that will be elaborated further when the deliverable will be submitted:

Lack of governance and regulation to support collaborative practices. Example: regulation to allow easily to appoint authorization: stakeholder B to collect a parcel if a parcel shipped to stakeholder A cannot be collect by stakeholder A. The need to white label public click and collect points. Such recommendation is considered to end up with a distributed system to manage traffic, reduce conjunction and facilitate on city decarbonization.

4.7 Lowering barriers for the adoption of big data analytics

Presented by Mauricio Fadel (BigDataStack)

Introduction

Nowadays, news about AI and big data analytics being successfully implemented in very different areas such as transportation, health and retail have become the norm; these AI and big data applications offer valuable insights and projections for decision making as well as automatizing great part of business processes and bringing benefits to several fields. However, it is important to note that all of these great achievements are not breakthroughs in the science of AI or big data, but are implementations of these techniques to different use cases. These implementations are possible thanks to knowledge breakthroughs and technology advances that have already happened mostly more than a decade ago. This means that there is a huge opportunity for all organizations to benefit from AI and big data; because the knowledge is already out there and needs engineering and implementation efforts. For an organization that successfully implements AI and/or big data analytics, these benefits include reducing costs, providing useful insights and increasing productivity.

Challenges for Big Data Adoption

However, this adoption is not an easy task and has not always been successful due to the big challenges big data presents. To use big data, organizations must define and properly implement multiple processes, overcoming several challenges. We can separate these challenges in two types: technical and legal. The first technical challenge is data availability.

For most companies, getting the data they need is not a trivial problem. Many times, this is because they do not have enough data for training machine learning models that require great amounts of data, or other times because they do not know how to capture the data required in their day-to-day operations. Another common problem is that even when they have the data, these might not be correctly labelled for their use and to label it requires a great effort. The second challenge of this type is skills shortage. Demand for data scientists has increased an amazing 344% from 2013 and big data related positions have followed a similar trend. These skills are fundamental when developing big data and AI projects, and can make the difference between a failure and a success. Another important technical barrier for organizations is security and privacy of data. There are data breaches every day; records that are exposed online with and without intention, data that is stolen, ransomware infecting thousands of computers every day and more. According to the IT Governance, a company specialized in IT security, 7.28B records have been breached this year so far. Including medical records, banking information and Facebook account details. Companies usually focus on their products and not on the security and privacy of the data they manage, making most companies insecure.

Moreover, organizations also encounter legal barriers when managing big data. Just as with security and privacy, companies usually do not possess knowledge about the laws related to data protection. To make matters worse, it is also hard to find experts in this topic, because laws change according to the country that the organization works in and laws and regulations are also continuously evolving and changing.

GDPR has brought big changes, and it has made a favour to companies because it unifies the laws that European companies must follow, instead of adapting to each different country. However, it is still not well understood and several companies are now afraid of managing data because of the hefty fines that GDPR imposes.
In particular, these challenges are significantly harder for SMEs and small organizations due to budget and skills limitations, creating an unfair disadvantage for them. In the current set up, Big Data and AI benefits are mostly enjoyed by big companies with the power, and money to get involved in it.

How can we help?

As professionals in the field, there are several actions we can take to reduce the gap between big data savvy organizations and those that are total beginners. One of these actions is the creation of open data initiatives from different governments in the world. The availability of more data is crucial for organizations and citizens, empowering them to analyse and use these data for a plethora of applications.
Another important action is to create a conscience of what AI and Big Data are, in what problems and circumstances they can help, and even more importantly, in which cases they are not useful. Nowadays, we are facing a big hype of big data because of the great benefits that brings as we have stated before, but because of this, many organizations want to implement AI and big data solutions to problems that can be easily solved by using more traditional approaches. It is important to use AI and big data analytics only when they are useful, otherwise, projects fail and organizations and the people involved get frustrated. In the past, this has led to AI winters and if we do not act responsibly, we might come to another one soon.

Finally, the development of tools for big data management and exploitation is another action which has a great effect in reducing this gap. Thanks to the availability of these tools, organizations need less effort and skills to implement big data solutions in their activity. Our take on this is BigDataStack: a platform that offers easy definition, development and management of big data applications and services for analytics. The BigDataStack platform aims to provide a set of tools and services to simplify the development and management of big data applications, covering the full data pipeline. It provides valuable information about the data traffic in applications and bases infrastructure management decisions on data analytics of the system and applications running on it. With BigDataStack, we aim to help organizations enjoy the benefit of big data applications, reducing development efforts and increasing the applications performance at the same time.

Lessons learnt

For the development of BigDataStack, we have analysed the needs of big data applications and services, and we are currently working on the implementation of 3 use cases from the naval, insurance and retail industries. While doing so, we have learnt some lessons about the barriers and challenges that big data projects involve.

The first thing to note even though it might seem obvious, but many times is not well understood, is how requirements vary for different types of applications, stakeholders and organizations. The criticality of the requirements vary greatly. For example, if a requirement is violated, in a predictive maintenance application for ships it might mean the ship could sink, for an insurance company this would mean losing money and in the retail case, one requirement violation might mean a page will load slowly.

Secondly, the applications' performance largely depends on their deployment, but also on their current load. Most times, it is very hard to anticipate these things before the application is already functioning and because of this, complying with their requirements is not guaranteed.
Furthermore, we have noticed that most organizations do not know what data they need to apply big data in their domain. They usually have a notion, but they do not know exactly what data they need, and what they need it for. This is crucial for GDPR, since it states that we need to declare clearly what data is saved and what it will be used for. Something that can help a lot to this is data visualization. It is expensive, because it needs time to think what we need to visualize and why, and also to develop the actual visualization, but it can help a lot to define the problem we are dealing with and how we can solve it.
Finally, we have seen that it is particularly hard to know the right amount of resources that Big Data applications will need in their final deployment. This is especially important for SMEs that cannot overprovision resources because they do not have as many resources as large companies do. Many scaling solutions that aim to help with this problem have become popular, like the ones offered by Amazon Web Services and Microsoft Azure, but still with these auto-scaling solutions, the mapping between performance and requirements is not always correct and applications' requirements can be violated. This is something that we also aim to solve with our project BigDataStack.

4.8 Big Policy Canvas roadmap for future research directions in data-driven policymaking

Presented by Francesco Mureddu, Lisbon Council

Big Policy Canvas aims at renovating the public sector on a cross-border level by developing a Roadmap that will enable public administrations to improve their readiness with regard to the integration of Big Data for the achievement of informed, evidence-based policy making in highly important application fields. The project approach for delivering its Roadmap will be based upon the following four major streams of activities: community building and networking, knowledge collection and analysis, Big Policy Canvas' innovation offering and road-mapping, and awareness creation and sustainability. The project has three main outputs:

- Big Policy Canvas Knowledge Base, is a state-of-the-art, online and dynamic repository that functions as an accumulator uniting all the knowledge produced during the project. It is structured along the three dimensions of needs, trends and assets and furthermore offers a mapping among them by defining how they are interconnected and how they influence each other;

- Big Policy Canvas Assessment Framework, is a methodology for mapping needs and trends to application domains, as well as for assessing the former in terms of their criticality or intensity respectively, with the ultimate goal of prioritising application domains and bringing forward those of greater interest, importance, urgency and capability for innovation;

- Big Policy Canvas Roadmap for Future Directions. The roadmap capitalises on the project outputs to carve the way for the future of EU policy-making and modelling.

Specifically, the aim of the Big Policy Canvas Roadmap for Future Research Directions in Data-Driven Policy Making is to put forward the different research and innovation directions that should be followed in order to reach the anticipated vision for making the public sector a key player in tackling societal challenges through new data-driven policy-making approaches. The road-mapping exercise builds on previous projects such as SONNETS, CROSSOVER, CROSSROAD, eGovRTD2020 and PHS2020, which adopted a policy-oriented approach including a foresight element by combining road-mapping with scenario building techniques. The road-mapping exercise encompasses three main steps:

1. Identification of the gaps that hinder the rapid and effective uptake of data-driven policy-making and policy-implementation solutions and approaches;

2. Elaboration of a set of future research challenges and application scenarios related to the use of big data in policy making;

3. Definition of a set of practical research directions and recommendations for all stakeholders involved.

The Big Policy Canvas roadmap has identified six main categories of research challenges:

- Privacy, transparency and trust. Even more than with traditional IT architectures, Big Data requires systems for determining and maintaining data ownership, data definitions, and data flows, especially when dealing with pervasive collection. Ubiquity of algorithms in everyday lives is an important reason to focus on addressing challenges associated with the design and technical aspects of algorithms and preventing bias from the onset. In fact, the use of algorithms for automated decision-making about individuals can result in harmful discrimination, and biased decision making (based on bias in the training data). Finally, opening big government data can increase transparency in policy making;

- Public Governance Framework for Data Driven Policy Making Structures. The governance notion stands for setting and managing rules guiding policy-making and policy implementation. Within the scope of electronic governance, evidence-based and data informed policy-making applies technology in order to efficiently transform governments, their processing of information and decision-making, and their interactions with citizens and businesses. In this framework, governance has to focus on how to leverage data for more efficient, rational, participative and transparent policy making, and specifically in helping understanding the problems that need to be addressed, considering potential alternatives and the ability to identify the best solution;

-Data acquisition, cleaning and storing. The appropriateness of any Big Data source for decision-making should be made clear to users, investigating systematic errors and the issue of representativeness. Any known limitations of the data accuracy, sources, and bias should be readily available, along with recommendations about the kinds of decision-making the data can and cannot support;

- Data clustering, integration and fusion. Combination and meaning extraction of big data stemming from different data sources to be repurposed for another goal. This requires the composition of teams that combine to types of expertise: data scientists, which can combine different datasets and apply novel statistical techniques; domain experts, that help know the history of how data were collected and can help in the interpretation. Interesting aspect is the need to ensure interoperability and exchange of data and information from different databases within the public administration;

- Modelling and analysis with big data. Here the point is to develop effective infrastructures merging the science of data with the development of highly predictive models, to come up with engaging and meaningful visualizations and friendly scenario simulation engines. Understanding the present through data is often not enough and the impact of specific decisions and solutions can be correctly assessed only when projected into the future. Hence the need of tools allowing for a realistic forecast of how a change in the current conditions will affect and modify the future scenario: scenario simulators and decision support tools;

- Data visualization. Making sense and extract meaning of data can be achieved by placing them in a visual context: patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software. This is clearly important in a policy making context, in particular when considering the problem setting phase of the policy cycle and the visualization often results of big data modelling and analysis.

The Big Policy Canvas Assessment Framework, Roadmap for Future Directions and Knowledge Base will be the foundation of further projects in the field, which will be able to go deeper in investigating the research challenges and in expanding the knowledge base.

4.9 Legal implementation barriers of privacy preserving technologies

Presented by Karolina La Fors, E-Sides

"Data-driven innovation is deeply transforming society and the economy. Although there are potentially enormous economic and social benefits this innovation also brings new challenges for individual and collective privacy, security, as well as democracy and participation. The main objective of the CSA e-SIDES is to complement the research on privacy-preserving big data technologies, by analysing, mapping and clearly identifying the main societal and ethical challenges emerging from the adoption of big data technologies, conforming to the principles of responsible research and innovation; setting up and organizing a sustainable dialogue between industry, research and social actors, as well as networking with the main Research and Innovation Actions and Large Scale Pilots and other framework program projects interested in these issues. It investigates stakeholders' concerns, and collect their input, framing these results in a clear conceptual framework showing the potential trade-offs between conflicting needs and providing a basis to validate privacy-preserving technologies. It prepares and widely disseminates community shared conclusions and recommendations highlighting the best way to ultimately build confidence of citizens and businesses towards big data and the data economy.[28]"

28 Bachlechner D.; Friedewald M.; Weitkamp, J.; Prill M. A.; La Fors, K.; Sears A. M.; Custers, B. ://e-sides.eu/resources/deliverable-42-overview-of-design-requirements

Legal barriers to the implementation of privacy-preserving technologies[29]

We identified four legal barriers applicable to innovating privacy-preserving technologies in diverse big data contexts:

1. EU-US regional differences: most influential big data technology companies are still US-based whereas currently the strongest data protection prescriptions are offered by the EU. Although privacy is codified as a constitutional right in the US, it is mainly approached from the perspective after harm occurs. In the EU privacy and data protection are fundamental rights that require protection as a starting point[30].

2. Sensitive data was identified, because violating the protection of this data type is perceived causing most significant harm for data subjects in the widest contexts of big data.

3. Inferred data was identified, because this type of data is not guaranteed by the EU data protection regime. Rendering data available for correlations is the underlying logic of all big data contexts in order to draw 'better' algorithmically facilitated decisions on individuals. However, this stretches the concept of identifiability of (personal) data to an extent where predicting privacy violations becomes close to impossible. Consequently, this in itself hampers the implementation of privacy-preserving technologies.

4. Liability and ethical responsibility was identified because whether or not and how privacy-preserving technologies are implemented have mutual implications in terms of liability and responsibility on all involved stakeholders that are part of interactions through big data.

29 This contribution is based on Deliverable 4.2. of the E-Sides project.

30 Moerel, L. & Prins, C, (2016) "Privacy for the Homo Digitalis: Proposal for a New Regulatory Framework for Data Protection in the Light of Big Data and the Internet of Things" Retrieved on 28thof April 2019 from https://ssrn.com/abstract=2784123

29 This contribution is based on Deliverable 4.2. of the E-Sides project.

30 Moerel, L. & Prins, C, (2016) "Privacy for the Homo Digitalis: Proposal for a New Regulatory Framework for Data Protection in the Light of Big Data and the Internet of Things" Retrieved on 28thof April 2019 from https://ssrn.com/abstract=2784123

Policy recommendations for addressing legal implementation barriers

1. The flexible interpretation of privacy and privacy-preserving technologies, which is both a blessing and a curse for professionals designing and using these technologies could be addressed by policies that offer guidelines how to integrate legal definitions of privacy into design requirements that are tailored to different big data contexts.

2. Policies aimed at bridging differences in EU and US approaches to privacy and competition law could help deconstruct implementation barriers for privacy-preserving technologies. Although US companies handling data of EU residents must comply with GDPR and align US and EU approaches to the right to data protection, the US approach remains quite different. Also, because competition law in the US is viewed to legitimise the use of consumer data "as a key competitive strategy" for companies which is not valid for the EU. Policies that could incentivize harmonization of EU-US interests could also deconstruct legal barriers for privacy-preserving technologies. This could include as a common denominator (both for public and private stakeholders) incentives centred around assessing and increasing the reliability and security of data input as well as that of the AI-mediated data flows by minimizing unlawful intrusion, (ab)use and errors.

3. Sector specific policies and best practices for the handling of sensitive data are also perceived as assets by a wide spectrum of professionals. This is so, because on one side of the spectrum certain healthcare professionals see strict privacy preservation as impeding epidemiological research, and on the other side others averse to share data due to the for them unforeseeable privacy implications on patients.

4.10 Overview of key data-related legislative frameworks

Presented by Julien Debussche, Bird & Bird

A presentation on the outcomes of the LeMo[31] project was provided, which covered legal Issues surrounding Data & Transport. The team has performed a legal scan around big data looking at the entire legal landscape, which resulted in the following overview of 13 legal topics:

Figure 3: Main components of the business pilot

The main point is that the legal landscape around data, and the connected policy issues, is a complex setup that involves many types of legislation, not only the (much discussed) GDPR. However, data privacy was one of the six topics that was further elaborated on:

31 lemo-h2020.eu

Privacy & Data Protection

Certain principles and requirements can be difficult to fit with some of the main characteristics of big data. Need of a balance between the various interests at stake. Must keep in mind that the right to the protection of personal data is not an absolute right, but must be considered in relation to its function in society and be balanced against other fundamental rights, in accordance with the principle of proportionality. Any guidance or decision should carefully take into account all interests at stake. Otherwise it would impede the development of disruptive technologies and prohibit the emergence of a true data economy.

Open Data

The 'big data' required to feed big data analytics tools emanates from different sources. One source is the public sector: it has been opening up certain of its datasets to the public. The EU has taken both legislative and non-legislative measures to encourage the uptake of open data. Especially the PSI Directive 2003/98/EC. Open data regimes also encounter a number of challenges (technical, economic and legal) not to be ignored. The proposal for a recast of the PSI Directive aims to address some of these concerns.

Free Flow of Data

Free flow of data is an ideal scenario where there are no (legal) barriers to cross-border data flows. New EU Regulation on the free flow of non-personal data eliminates restrictions to cross-border data flows and their negative impact on business. Companies expect cost reductions (through elimination of data localisation requirements), expected to increase competition (of the EU cloud services market), start-ups expected to go to market quicker, increase innovation pace, improve scalability and achieve economies of scale. But uncertainties remain, including a difficult interaction with the GDPR.

Data Sharing Obligations

Examines legal instruments that impose specific data sharing obligations on private undertakings and therefore affect a company's control of, access to, or use of data. Usually sector-focused instruments and provide for an array of rights and obligations in relation to specific types of data in particular circumstances. Data sharing obligations are increasingly adopted in certain sectors and contexts (e.g. of Intelligent Transport Systems). Need to carefully consider whether the imposition of such general data sharing obligations is in each case equally necessary.

Data Ownership

Numerous stakeholders in the (big) data analytics lifecycle cannot rely on general exclusive rights (mainly intellectual property rights). Stakeholders increasingly try to claim "ownership" in (parts of) the datasets, BUT no specific ownership right in data the existing data-related rights do not respond sufficiently or adequately to the needs of the actors in the data value cycle. The only current solution is capturing the relationships between the various actors in contractual arrangements. However filling the data ownership gap with contractual arrangements is far from ideal.

Data Sharing Agreements (DSA)

Provide analysis of the common practice to use DSA to govern the access to and/or exchange of data between stakeholders in a big data lifecycle. Unclear whether DSA enables covering all possible situations with the necessary legal certainty. DSA entails numerous limitations in the absence of a comprehensive legal framework regulating numerous rights (e.g. ownership, access or exploitation rights) attached to data, the way in which such rights can be exercised, and by whom. Guidance has been issued by the EU Commission BUT a more solid and legally secure solution might be desirable.

Common Dissemination Booster

BIG DATA VALUE
PUBLIC-PRIVATE PARTNERSHIP

Common Dissemination Booster