D5.2 Roadmap for Future Research Directions
(Pending European Commission Approval)

Document name:	D5.2 Roadmap for Future Research Directions (Pending European Commission Approval)					**Page:**	1 of 127
Reference:	D5.2	**Dissemination:**	RE	**Version:**	1.0	**Status:**	Submitted

Document Information

List of Contributors	
Name	Partner
Francesco Mureddu	Lisbon Council
David Osimo	Lisbon Council
Esther Garrido	ATOS
Ricard Munné	ATOS
Juliane Schmeling	FOKUS
Vittorio Loreto	Sony Computer Science Laboratories
Peter Parycek	Danube University Krems
Gianluca Misuraca	JRC Seville
Giuseppe Veltri	University of Trento

List of Most Relevant Commenters	
Name	Institution
Akrivi Vivian Kiousi	Senior Business Development Manager, Head of Transport Lab Research and Innovation at Intrasoft
Alan Hartman	Senior Lecturer at Department of Information Systems, University of Haifa
Angela Guarino	Policy Officer at the European Commission
Anna Triantafillou	Deputy Head of Innovation Lab at Athens Technology Center
Basanta Thapa	Researcher at Fraunhofer FOKUS
Carlos Agostinho	Director of Operations at Knowledgebiz
Christos Botsikas	Information Technologist at National Technical University of Athens
Enrico Ferro	Head of Innovation Development Department at Links Foundation
Evmorfia Biliri	Researcher at Fraunhofer FOKUS
Gianluca Misuraca	Senior Scientist at JRC Seville
Giuseppe Veltri	Professor at University of Trento
Juliane Schmeling	Researcher at Fraunhofer FOKUS
Luca Alessandro Remotti	Business Innovation Project Manager at Join Institute of Innovation Policy
Maria Wimmer	Professor at University of Koblenz-Landau
Mariam El Ouirdi	Researcher at the University of Antwerp
Shefali Virkar	Research Associate at Donau-Universität Krems
Spiros Mouzakitis	Researcher at National Technical University of Athens
Vittorio Loreto	Director at Sony Computer Science Laboratories
Yannis Charalabidis	Professor at University of Aegean, World's 100 Most Influential People in Digital Government

Table of Contents

Document Information 2
Table of Contents 4
List of Tables 6
List of Figures 7
List of Acronyms 8
Executive Summary 9
1 Introduction 11
1.1 Purpose of the document 11
1.2 Relation to other project work 11
1.3 Structure of the document 12
2 Methodology 13
2.1 Roadmapping Exercise 13
2.2 Methodology for Gap Analysis 14
2.2.1 Process for Gap Identification 15
2.3 Input Collection Activities Performed 17
2.3.1 Input from Experts as a Memo 17
2.3.2 Input at the Big Data Value Forum in Vienna 18
2.3.3 Input from Experts on the Roadmap Structure in Commentable Format 19
2.3.4 Input from other Events 24
3 Current Status and What is New 26
3.1 The Policy Cycles 26
3.2 The Traditional Tools of Policy Making 28
3.3 The Key Challenges of the Policy Makers 29
3.4 Big Data Driven Policy Making 32
3.4.1 Big Data Value Chain 32
3.4.2 Big Data in the Policy Cycle 36
3.4.3 Bottlenecks and Enablers of Data-Driven Policy Making 38
4 Identification of Gaps and Research Needs 44
4.1 Step 1: Needs Selection 44
4.2 Step 1 to Step 4: Need Breakdown, Asset Assessment and Gap Identification 44
5 Research challenges on the use of big data for policy making 52
5.1 Research Clusters 52
5.1.1 Cluster 1- Privacy, Transparency and Trust 52

5.1.2 Cluster 2 - Public Governance Framework for Data Driven Policy Making Structures 54
5.1.3 Cluster 3 - Data Acquisition, Cleaning and Representativeness 54
5.1.4 Cluster 4 - Data Storage, Clustering, and Integration 55
5.1.5 Cluster 5 - Modelling and Analysis with Big Data 56
5.1.6 Cluster 6 - Data Visualization 57
5.2 From Research Gap to Research Clusters 59
5.3 Research Challenges 60
5.3.1 Research Challenges on Privacy, Transparency and Trust 61
5.3.2 Research Challenges on Public Governance Framework for Data Driven Policy Making Structures 74
5.3.3 Research Challenges on Data acquisition, Cleaning and Representativeness 79
5.3.4 Research Challenges on data storage, clustering, and integration 85
5.3.5 Research Challenges on Modelling and Analysis with Big Data 87
5.3.6 Research Challenges on Data Visualization 97
Conclusion 105
References 106
6 Annex I - Assets assessment against Needs functionalities (Step 3) 116

List of Tables

Table 1 – Research Clusters and Research Challenges 9
Table 2 - Table for assessment of assets against Needs functionalities 16
Table 3 - Needs selected for Gap analysis 44
Table 4 - Gap identification for N-S-1, Development of domain specific target and indicator systems 44
Table 5 - Gap identification for N-S-2, Involvement of the public and citizens, as well as the development of citizen-centred policy making 45
Table 6 - Gap identification for N-S-4, Strengthen citizens' trust in public administration 46
Table 7 - Gap identification for N-S-9, Cross-linked information exchange 47
Table 8 - Gap identification for N-O-7, Standardisation of processes 47
Table 9 - Gap identification for N-T-1, Cope with the production of huge volumes of data 48
Table 10 - Gap identification for N-T-3, Ensuring data security taking into account the protection of citizens' privacy 48
Table 11 - Gap identification for N-T-4, Establishment of a comprehensive technical infrastructure and IT architecture 49
Table 12 - Gap identification for N-I-1, Link between impact, quality, performance measurements and financial information 49
Table 13 - Gap identification for N-I-3, Ensure availability of (real-time) information and knowledge 50
Table 14 - Gap identification for N-I-4, Comprehensive knowledge and information management 51
Table 15 – Mapping of research needs and clusters of research challenges 59
Table 16 – Research clusters and related research challenges 60
Table 17 – Taxonomy of Open Government Data research areas and topics 71
Table 18 - Validation Procedures for Simulation Models 89
Table 19 - Asset assessment for N-S-1 116
Table 20 - Asset assessment for N-S-2 116
Table 21 - Asset assessment for N-S-4 117
Table 22 - Asset assessment for N-S-9 119
Table 23 - Asset assessment for N-O-7 120
Table 24 - Asset assessment for N-T-1 120
Table 25 - Asset assessment for N-T-3 121
Table 26 - Asset assessment for N-T-4 122
Table 27 - Asset assessment for N-I-1 122
Table 28 - Asset assessment for N-I-3 - Use Case 123
Table 29 - Asset assessment for N-I-3 - Code list / Ontology / Taxonomy / Vocabulary/Standard 124
Table 30 - Asset assessment for N-I-3 - Application 124
Table 31 - Asset assessment for N-I-3 - Tool 125
Table 32 - Asset assessment for N-I-3 - Portal/Database/Data source 126
Table 33 - Asset assessment for N-I-3 - Model 127
Table 34 - Asset assessment for N-I-4 127

List of Figures

Figure 1 – Methodology for the Elaboration of the Roadmap 14
Figure 2 - Methodology for gap identification 16
Figure 3 – Screenshot from the Roadmap Structure in Commentable Format 20
Figure 4 – Display of Comments in the MakingSpeechTalk tool 21
Figure 5 - Most used words in speech (left) and comments (right) 21
Figure 6 - Most used words in speech (left) and comments (right) 22
Figure 7 – Policy Cycle and Related Big Data Activities 27
Figure 8 - The big data-revised policy cycle 28
Figure 9 – Small and Big Data 33
Figure 10 – Big Data Value Chain and Technologies 34
Figure 11 – Digital Data Market 35
Figure 12 – Structure of the Research Clusters 52
Figure 13 - Taxonomy of objections to algorithmic decision-making 63
Figure 14 – Automation of Government Services (Source: Engin and Treleaven 2019) 64
Figure 15 - Six types of ethical concerns raised by algorithms (Source: Mittelstadt et al. (2016) 66
Figure 16 - Key research themes in response to barriers to legitimate and effective algorithmic governance (Source: Danaher et al. 2017) 67
Figure 17 – Artificial intelligence impact assessment canvas (Source: Engin and Koshiyama 2019) 68
Figure 18 – Framework prototype (Source: Tal et al. 2019) 68
Figure 13 – Data Modelling Approach with Machine Learning 88

List of Acronyms

Abbreviation / acronym	Description
AGCOM	Autorità per le Garanzie nelle Comunicazioni
ADMS	Asset Description Metadata Schema
AI	Artificial Intelligence
BDTI	Big Data Test Infrastructure
BPC	Big Policy Canvas
BDVA	Big Data Value Association
CEF	Connecting Europe Facility
DAF	Data & Analytics Framework
DCAT	Data Catalogue Vocabulary
Dx.y	Deliverable number y belonging to WP x
EC	European Commission
EU	European Union
GDPR	General Data Protection Regulation
ICT	Information and Communication Technologies
IDC	International Data Corporation
IT	Information Technology
KB	knowledge base
NLP	Natural Language Processing
RC	Research Cluster
PA	Public Administration
WP	Work Package

Executive Summary

This document presents the second version of the Big Policy Canvas Roadmap for Future Research Directions in Data-Driven Policy Making, which aims 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. Specifically, after the introduction, the document presents the methodology for the elaboration of the roadmap, highlighting in particular the crowdsourcing activities which are ongoing. Then the document presents an introduction of the state of play in the use of Big Data in policy makings, highlighting what kind of challenges this technology can help to cope with. Further, the document discusses a set of gaps and research needs in the use of Big Data in policy making. Based on the gaps and needs we define six main research clusters related to the use of Big Data in policy making. Four of them are built on the Big Data cycle and value chain, while two are transversal at each phase of the cycle. For each research cluster, we define and briefly present a set of research challenges (see Table 1).

Table 1 – Research Clusters and Research Challenges

Research Cluster	Research Challenges
C1- Privacy, Transparency and Trust	RC 1.1 - Big Data nudging
	RC 1.2 - Algorithmic bias and transparency
	RC 1.3 - Open Government Datasets
	RC 1.4 – Manipulation of statements and misinformation
C2 - Public Governance Framework for Data Driven Policy Making Structures	RC 2.1 - Forming of societal and political will
	RC 2.2 - Stakeholder/Data-producer-oriented Governance approaches
	RC 2.3 - Governance administrative levels and jurisdictional silos
	RC 2.4 - Education and personnel development in data sciences
C3 - Data acquisition, cleaning and representativeness	RC 3.1 – Real time big data collection and production
	RC 3.2 - Quality assessment, data cleaning and formatting
	RC 3.3 - Representativeness of data collected
C4 - Data storage, clustering and integration	RC 4.1 - Big Data storage and processing
	RC 4.2 - Identification of patterns, trends and relevant observables
	RC 4.3 - Extraction of relevant information and feature extraction
C5 - Modelling and analysis with big data	RC 5.1 - Identification, acceptance and validation of suitable modelling schemes inferred from existing data
	RC 5.2 - Collaborative model simulations and scenarios generation
	RC 5.3 - Integration and re-use of modelling schemes
C6 - Data visualization	RC 6.1 – Automated visualization of dynamic data in real time
	RC 6.2 - Interactive data visualization

Finally, the document presents the next step for ensuring the sustainability of the work carried out within the scope of the roadmap, namely the elaboration of a joint JRC-BDVA Scientific Report building on the roadmap, to be co-authored by Francesco Mureddu (Lisbon Council), Juliane Schmeling (FOKUS), Gianluca Misuraca (Senior Scientist at the JRC Seville and member of the expert committee), and the Big Data Value Association Smart Cities sub-group. The Scientific Report will be first presented at the High-Level Conference on Data Economy, taking place in Helsinki, on 25-26 November 2019.

1 Introduction

1.1 Purpose of the document

It can be taken for granted that technological advancements, especially in the last decade, have revolutionized the way that both every day and complex activities are conducted. It is indicative that practically all expectations regarding innovation, regardless of the domain of application, are most of the time directly linked with the exploitation of emerging technologies, as well as with the constantly increasing volume of available data. It is, thus, expected that a particularly important actor, such as the public sector, should constitute a successful disruption paradigm through the adoption of novel approaches and state-of-the-art ICTs. New concepts, especially those that consider the available data as a way of ensuring accurate and meaningful input to public sector organisations that can help establish new types of evidence-informed policies, are of the utmost importance. However, despite the investments continuously performed and initiatives implemented in the field of public sector modernisation, it is really hard to allege that "we are already there" when it comes to full exploitation of ICT innovations and data towards aiding the public sector to meet the societal and financial challenges that are emerging. Big Policy Canvas aims at renovating the public sector on a cross-border level by mapping the needs of public administrations with methods, technologies, tools and applications from both the public & the private sector, stepping upon the power of open innovation and the rich opportunities for analysis and informed policy making generated by big data. As a result, the project will deliver a live roadmap that will propose short and midterm milestones and relevant actions needed towards achieving the expected impacts for the public sector and the society at large. The consolidation of such a roadmap, as envisioned by Big Policy Canvas will be based upon a highly collaborative and multidisciplinary approach and will take into account both completed and ongoing similar activities within and outside the European Union. 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.

1.2 Relation to other project work

The deliverable 5.1 is the outcome of Tasks 5.1 and 5.2 of the project. Furthermore, the gaps and research needs will be identified by comparing the needs of public administrations identified in WP3 (D3.3 Needs and Trends Assessment with a multidisciplinary Big Data perspective) and the potential to be covered through the exploitation of existing technologies Methods, Tools, and Applications in WP4 (D4.2 Methods, Tools, Technologies and Applications). The gap analysis and the identification of research needs will form the basis for the identification of research challenges to be included in the roadmap. Regarding the chapter on research challenges, the analysis is primarily conducted by the project

consortium, but as reported in the methodology chapter it has also gained from the interaction with the stakeholders' community in WP6 and from the input from the experts contracted in the project. D5.1 represents the first version of the roadmap. The second version, D5.2, will be released at M24.

1.3 Structure of the document

The rest of this document is structured as follows:

- **Chapter 2** reports on the methodology currently followed to produce the roadmap;
- **Chapter 3** reports on the current status and what is new about the use of big data for policy making;
- **Chapter 4** presents the gap analysis, stemming from the triangulation of needs, trends and assets;
- **Chapter 5** presents the preliminary research challenges. Research challenges are currently grouped in 5 clusters, and for each research challenge we present a description and link with the state of the art, current status, importance in the policy making process, and application cases;
- **Chapter 6** finally reports next steps.

2 Methodology

2.1 Roadmapping Exercise

The roadmapping exercise will build on previous projects such as SONNETS[1], CROSSOVER[2], CROSSROAD[3], eGovRTD2020[4] and PHS2020[5], which adopted a policy oriented approach including a foresight element by combining roadmapping with scenario building techniques. In this context, the project's roadmapping approach will follow such a holistic approach. In particular, this pillar contains all the activities required to provide future research directions for the public administrations by simultaneously providing policy, research and industry recommendations with a view towards the EC's strategy for H2020 and beyond. Results from the previous project activity will be combined appropriately to reveal all aspects that should be tackled in the future research activities and those that could be pushed in the short or long term for exploitation. Recorded methods, technologies and tools will be mapped to the public sector needs that they address, the trends they exploit and the possible challenges/barriers they either meet or overcome, in order to come up with an evidence-based gap analysis in the public administration sector. These actions will allow the definition of implementation as well as of research challenges and their transformation into recommendations for the next work programmes.
The roadmapping 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.

Clearly, the core activity of the roadmap lies on the elaboration of the research challenges and policy recommendations, and in particular on the answers to the following questions:

- Which major research challenges should be considered and addressed for evidence-based policy-making and policy-monitoring in order to tackle wider societal challenges
- Which sub-challenges and emerging technologies are part of these major challenges?
- How are these challenges related to the different policy domains?
- What kind of instruments are necessary to tackle these challenges?
- What is the anticipated impact of these challenges to each policy domain and to the society?
- Which are the broad recommendations for policy makers, researchers and industry that are meaningful to accelerate the roadmap's take-up?

The overall methodology for the elaboration of the roadmap is presented in the chart below, in which blue boxes represent portions of the activities already carried out, the purple boxes/arrows represent input to the roadmap, while golden arrows represent advancements in the production of the roadmap.

[1] Info on the project available at https://www.sonnets-project.eu
[2] Info on the project available at https://cordis.europa.eu/project/rcn/100802/factsheet/en
[3] Info on the project available at https://cordis.europa.eu/project/rcn/93842/factsheet/en
[4] Info on the project available at https://cordis.europa.eu/project/rcn/79311/factsheet/en
[5] Info on the project available at https://cordis.europa.eu/project/rcn/85297/factsheet/en

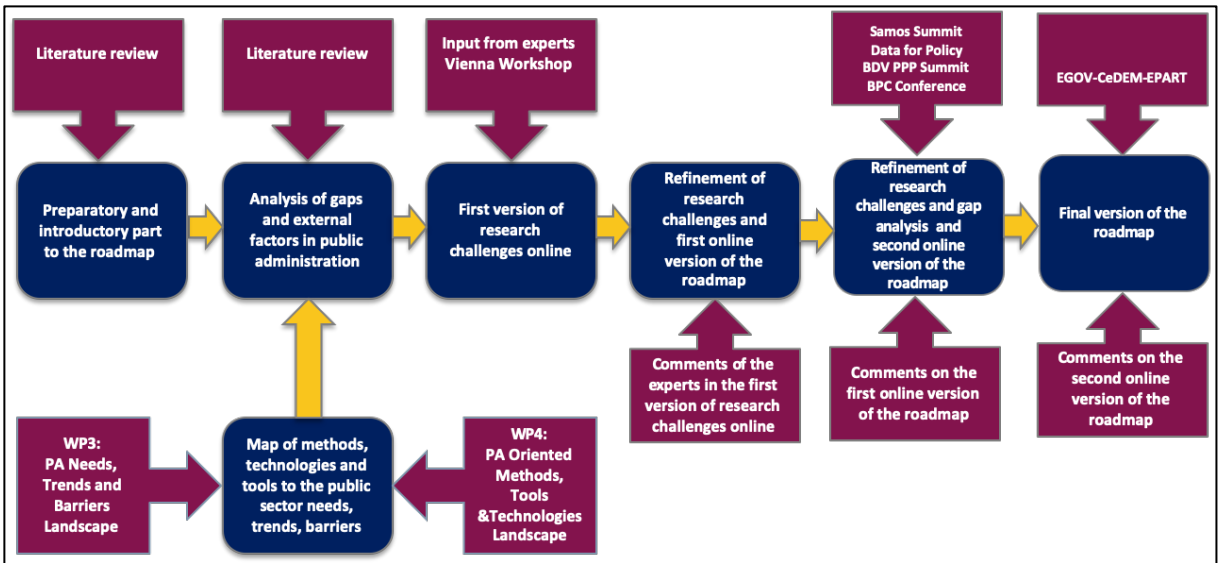

Figure 1 – Methodology for the Elaboration of the Roadmap

As it can be seen, a first literature review has been necessary to collect relevant papers and information to produce the preparatory part and the introduction to the roadmap (chapter 3). As a second step, the output produced in WP4, i.e. the mapping of methods, technologies and tools with respect to the public sector needs, trends and barriers, has been used to produce an analysis of gaps and external factors hindering the adoption of big data technologies in the public sector (chapter 4, and more on that in §2.2). The first version of research clusters and challenges has been elaborated building on the preparatory phase, the analysis of gaps and external factors, the direct input from experts provided via a set of memos and the input provided in the second Big Policy Canvas workshop in Vienna. The first version of research clusters and challenges has been uploaded in MakingSpeechTalk, a proprietary software of the Lisbon Council and has gained substantial input (e.g. direct comments on the tool) provided by the experts and their team. Then the consortium has developed a first version of the roadmap depicted in deliverable D5.1 This version has been presented in several events/workshops, and a synthesis has also been uploaded in MakingSpeechTalk for comments. A second version of the roadmap has then been uploaded for comments, and also presented in a last event in September 2019.

2.2 Methodology for Gap Analysis

The Big Policy Canvas Gap analysis is focusing on the gaps that hinder the rapid and effective uptake of data-driven policy-making, policy-modelling and policy-implementation solutions and approaches. The gaps are identified by comparing the needs of public administrations identified in WP3 and the potential to be covered through the exploitation of existing Methods, Tools, Technologies and Applications, that is, the assets, identified in WP4.

Gaps are the mismatch between what currently can be provided in this universe through the use of the existing assets, and what are the current needs in terms of information, organisation, strategy, legal and technology according the current conceptual, societal and technological trends.

We look for gaps in existing assets to fulfil the needs and advance in the policy making through the use of (big) data for the evidence-based policy-making. Firstly, we need to identify the gaps from the needs collected in WP3, D3.3, and the assets collected in WP4, D4.2. Needs require some functionalities to be fulfilled, so we have identified those functionalities and subsequently to which extent the associated assets support those functionalities according the trends that foment those needs.

2.2.1 Process for Gap Identification

For the identification of the gaps we follow the methodology described below. The main inputs for the gaps identification are the description of need and Assets found in the Big Policy Canvas Knowledge Base.

Step 1: Selection of needs from the KB[6] is based on those with high priority, according the prioritisation performed in the assessment framework, updated in D3.3, and those with big data potential, according to the big data potential relevance assessed as well in D3.3, as the project is focused in the policy development through the use of (big) data evidence.

Step 2: After the needs are selected, each one is broken down into the functionalities that form that need. For example, form the description of need N-S-1, *Development of domain specific target and indicator systems*, two main functionalities that are required from the Need description:
"Already the political economist and sociologist Max Weber once has pointed out that decision makers need to ensure the rationality of their decisions, by trying to balance out the best relation of means and ends.
Consequently, policy makers need to clarify the targets that they want to reach through certain political programmes and norms. In fact, the executive bodies need quite precise targets, since they are responsible for the adoption and implementation of political and legal solutions and need to translate political solutions in concrete activities. If public administrations want to monitor political targets, they need to set up a management control system, as it is already quite common in the private sector. Nevertheless, since it is not possible to score success from insulated financial ratios (See also Need: Link between impact, quality, performance measurements and financial information), the public sector needs to observe much more complex systems in consideration of public interests.
In a conducted interview with a public administration representative on the regional ministerial level in the youth welfare policy domain, the interviewee confirmed that there is a lack of clearly formulated goals on the political level. The interviewee further mentioned that without clear goals on a political level, executive bodies are incapable to derive operationalised goals and indicators.
A second problem he mentioned is that targets, if they are formulated, should be well balanced among each other, since it is important in the implementation phase to know which targets have priority to set up a strategic planning. For example, it is difficult to implement child day care availability for everybody and best trained childcare workers at the same time.
To sum up, policy domain specific targets and indicator systems are especially relevant in the formulation, and implementation phase, but are also relevant in the monitoring and evaluation phase, since it is impossible to monitor and evaluate political targets and their derived indicators in a performance measurement system without targets."

Error! Reference source not found. presents the gap identification methodology.

[6] Big Policy Canvas, Knowledge Base, https://www.bigpolicycanvas.eu/community/kb, retrieved March 2019

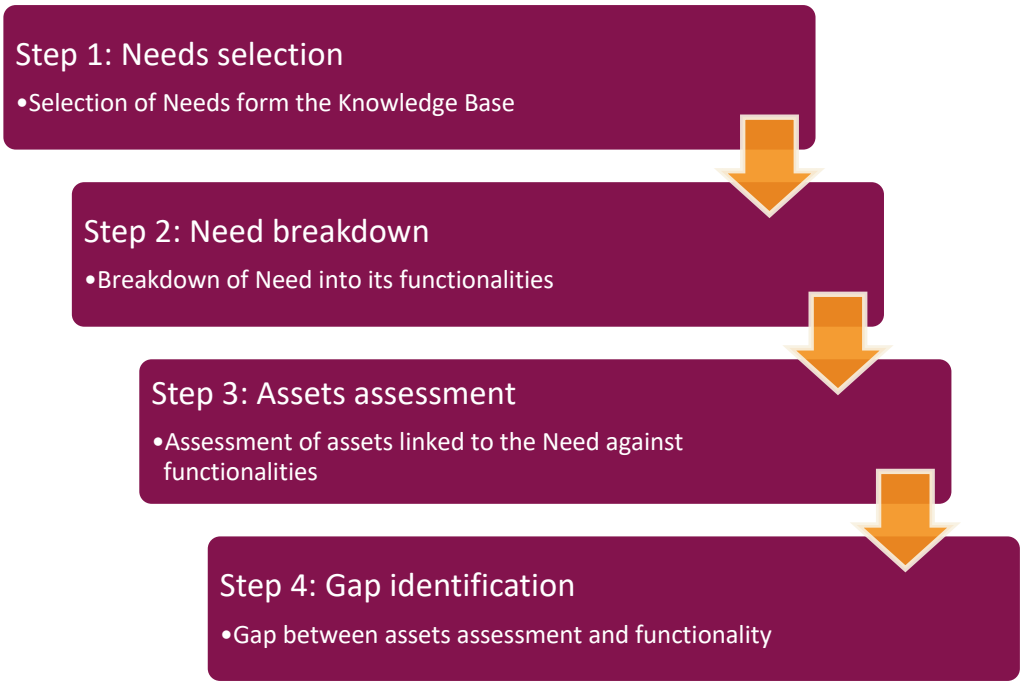

Figure 2 - Methodology for gap identification

So, from the two parts of the text highlighted above we have concluded that the following functionalities are required:

- F1: Management control system to monitor political targets based on multiple indicators (impact, quality, performance measurements and financial information).
- F2: Definition of clear goals in policy building with balanced targets.

Typically, one or two functionalities are found for each need.

Step 3: Once the needs functionalities are defined, we assess each asset linked to the need against the functionalities extracted. This is performed with a table with the functionalities in the *x* axis, and the assets in the *y* axis, as in the example in Table 2 below. At the end we have a set of assessments for each functionality from each asset.

Table 2 - Table for assessment of assets against Needs functionalities

Functionalities / **Assets**	**F1: Functionality 1**	**F2: Functionality 2**
Asset 1	Assessment Asset 1 against Functionality 1	Assessment Asset 1 against Functionality 2
Asset ..	Assessment Asset .. against Functionality 1	Assessment Asset .. against Functionality 2
Assent N	Assessment Asset 2 against Functionality 1	Assessment Asset 2 against Functionality 2

Step 4: The gap is then extracted from the maximum level of compliancy of the assets against a given functionality. The space between this level of compliancy and the requirement of the functionality is identified as the gap for that functionality, so at least one gap is identified for each functionality.
One issue found during the assets functionality compliance assessment is that the description in the Knowledge Base[7] in quite limited for many of them, so at the end it is required to go through the description of the asset found in the corresponding website linked to it. This poses a homogeneity issue, as usually there is not the same level of detail, or even the same approach in the assets' description.

2.3 Input Collection Activities Performed

The elaboration of the roadmap has gained from several inputs provided by experts and attendants to events through several channels. Specifically, input has been provided directly from the experts contracted by the project under a form of a memo, by attendants to the Big Policy Canvas workshop celebrated within the scope of the Big Data Value Forum in Vienna, the Big Policy Canvas final conference in Venice, as well as from four main workshops: Data for Policy 2019 – Digital Trust and Personal Data (London, 11-12 June 2019); EGOV2019 – Joint conference EGOV-CeDEM-EPART 2019 (San Benedetto del Tronto, 2-5 September 2019); The 9th Samos 2019 Summit On ICT-enabled Governance in conjunction with The 6th International Summer School On Open and Collaborative Governance (Samos, 1-5 July 2019); and the BDV PPP Summit – Impact Empowered by Data-Driven Artificial Intelligence (Riga, June 26-28 2019). Finally, a great deal of input has arrived as comments on the roadmap in commentable format.

2.3.1 Input from Experts as a Memo

The first direct input to the roadmap was provided by the contracted experts of the project in December 2018. Such input, provided by four experts (Loreto, Misuraca, Peter Parycek, Veltri) has been the basis for the elaboration of the first structure of research challenges. Specifically, the input provided is as follows:

- Vittorio Loreto - *Notes on research challenges and application scenarios related to the use of big data in policy making.* This note focus on the use of big data for coping with unanticipated knowledge through the construction of scenario simulators and decision support tools. Specifically, he calls for the launch of new research directions aimed at developing 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. The research challenges and perspectives in clusters 5 and 6 stem from this contribution as well as from other discussions with Vittorio Loreto and other experts;
- Gianluca Misuraca - *How could policy making take advantage of current developments in data and, at the same time, address ethical concerns surrounding their use?* The note discusses the legitimacy of hyper-nudging in government, i.e. combination of predictive analytics with behavioral nudge applications. Big Data analytic nudges are extremely powerful due to their continuously updated, dynamic and pervasive nature, working through algorithmic analysis of

[7] Big Policy Canvas, Knowledge Base, https://www.bigpolicycanvas.eu/community/kb, retrieved March 2019

data streams from multiple sources offering predictive insights concerning habits, preferences and interests of targeted individuals;

- Peter Parycek – *Social Scoring through Big Data*. The note argues that as long as data is used for statistical issues like anonymous monitoring activities, early warning system for societal phenomenon or agent-based simulation, it could provide great value without limiting freedom or human rights. However, the situation becomes critical when governments and their administrations start to use big data for profiling purposes through combining personal small data with big data sources, e.g. grid investigation or nudging purposes. The research challenges and perspectives in clusters 1 and 2 stem from the contributions by Misuraca and Parycek, as well as from other discussions with those some experts and other experts;
- Giuseppe Veltri - *Emerging methodological aspects of using Big Data.* In the first part, the note focused on three main positive elements of big data, such as the fact that they increase size and resolution of the databases, they span across time (providing therefore time series) and they allow for non-reactive heterogeneity and behavioural consideration. In the second part, it is stressed that size of data alone cannot eliminate the potential presence of systematic errors, and that other methodological aspects of big data have to be carefully evaluated, such as the issue of representativeness and the construct validity problem. The research challenges and perspectives in cluster 3 and 4 stem from this contribution as well as from other discussions with Giuseppe Veltri and other experts;

2.3.2 Input at the Big Data Value Forum in Vienna

The 2nd Big Policy Canvas workshop was celebrated in the framework of the European Big Data Value Forum (EBDVF)[8], held in Vienna (Austria) on the 14th of November 2018. The objective of the second BPC Workshop was, on the one side, to get feedback on the research challenges prepared by BPC experts as the initial step for elaborating the BPC Roadmap for Future Research Directions in Data-Driven Policy Making and, on the other side, to announce and make the collaboration between BPC and BDVA effective, mobilising community around this topic. In the second and interactive part of the workshop, the WP5 leader presented an initial list of future research challenges building on the aforementioned content proposed by the BPC Experts' Committee. The objective of such interactive part was to elaborate on this initial list and to propose new research challenges and applications. More specifically, the WP5 leader clustered the research challenges presented them according to the topic, also considering the trends and needs analysis carried out in the project. He asked the audience to split into four different groups according to their interest in:

1. Transparency;
2. Data Ownership, Privacy and Security;
3. Data Collection and Linking;
4. Simulation and Decision Support Tools.

And then asked to answer three questions in the sheet of paper provided:

- Please add a research challenge related to your cluster
- Why is this challenge important in the policy making process? Which is the need that is addressed in the policy making process?
- Which technologies should be developed to cope with this challenge?

[8] https://www.european-big-data-value-forum.eu/

Each group was facilitated by one consortium member.
At the end of the session, the WP5 leader presented the input to the participants, who voted for the best research challenge proposed using a voting tool.
Apart from the valuable input gathered by the attendants, interesting input and discussions took place with two speakers:

- Mr. Raffaele Lillo, former Chief Information Officer of the Italian Digital Team, who presented the DAF: Data & Analytics Framework, which has the goal of improving and simplifying the interoperability and exchange of data between public administrations, promoting and improving the management and usage of Open Data, optimising activities of analysis and knowledge generation. The DAF optimise data exchange between public administrations and Open Data deployment, minimising transaction costs for data access and data usage and facilitating data analysis and data management by data scientist teams within the public administration;
- Mr. Mihkel Solvak, from the University of Tartu, gave Estonian perspective on how Estonian public administration is speeding up the policy cycle adding value to data. He presented X-Road, an open source data exchange layer solution that enables organisations to exchange information over the Internet. X-Road is a centrally managed distributed data exchange layer between information systems that provides a standardised and secure way to produce and consume services. X-Road ensures confidentiality, integrity and interoperability between data exchange parties

Both tools are presented as application cases in the research challenges below.

2.3.3 Input from Experts on the Roadmap Structure in Commentable Format

The first structure of the research challenges chapter was released at the end of February 2019 in commentable format by mean of a proprietary tool of the Lisbon Council called MakingSpeechTalk[9] (see Figure 3 and Figure 4).

[9] The roadmap structure in commentable format is available at https://roadmap.bigpolicycanvas.eu/ch/BPCRoadmap

BigPolicyCanvas

Big Policy Canvas Research Roadmap

HOME | STATISTICS | LOGIN

What are the Research Challenges for the use of Big Data in Policy Making?

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

At this stage, we present an initial set of research challenges categorised in **four main clusters.** By clicking on any texts below and you will be able to directly comment, line by line. Specifically, we kindly ask you to:
- Comment on the cluster classification (e.g. add others)
- Help refine existing research challenges
- Add research challenges that you deem important

For more information, please visit the project website or contact us at **francesco.mureddu@lisboncouncil.net**

Index

Research Clusters and Related Research Challenges
Cluster 1 – Privacy, transparency and trust
Cluster 2 – Data acquisition and cleaning and recording
Cluster 3 – Data clustering, integration and fusion
Cluster 4 – Modelling and analysis with Big Data
Cluster 5 – Data visualization

Figure 3 – Screenshot from the Roadmap Structure in Commentable Format

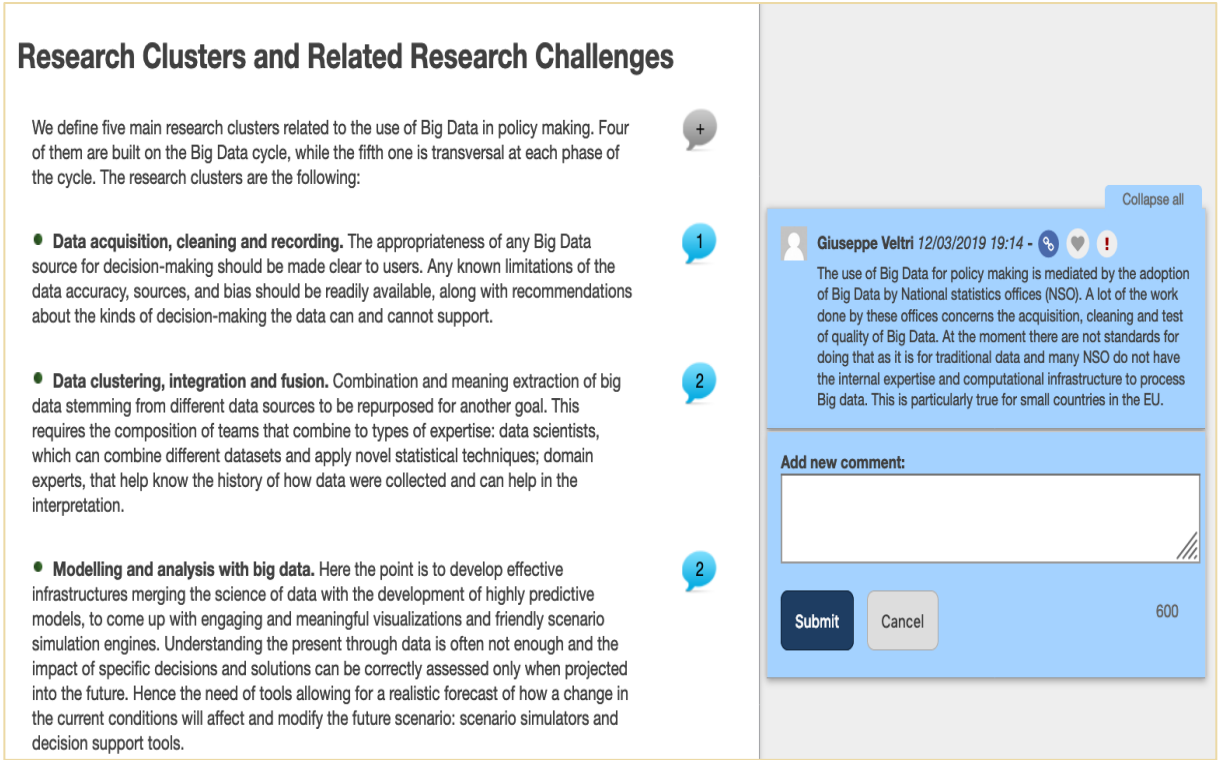

Figure 4 – Display of Comments in the MakingSpeechTalk tool

In this first iteration the tool collected 63 comments from 8 different commenters who provide 1546 words of comments. It has to be noticed that at this point the commenting was open only for the experts and their immediate teams. Some other statistics are available in Figure 5 below.

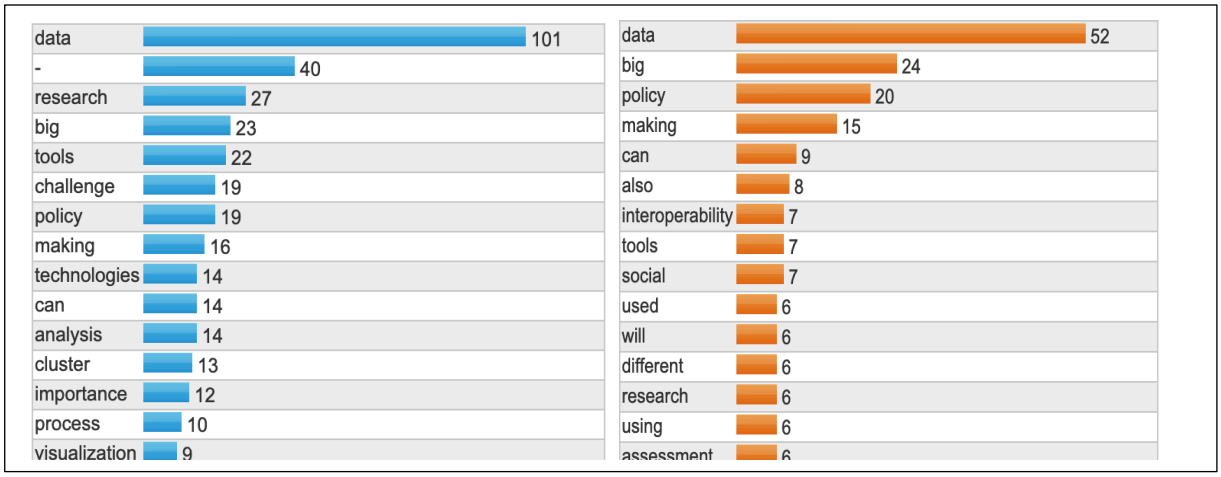

Figure 5 - Most used words in speech (left) and comments (right)

A first version of the roadmap, containing a synthesis of Chapter 3, 4 and 5, has been uploaded online in commentable format in May 2019, and has collected 268 comments from 38 commenters, for a total of 6.113 Comments words count.

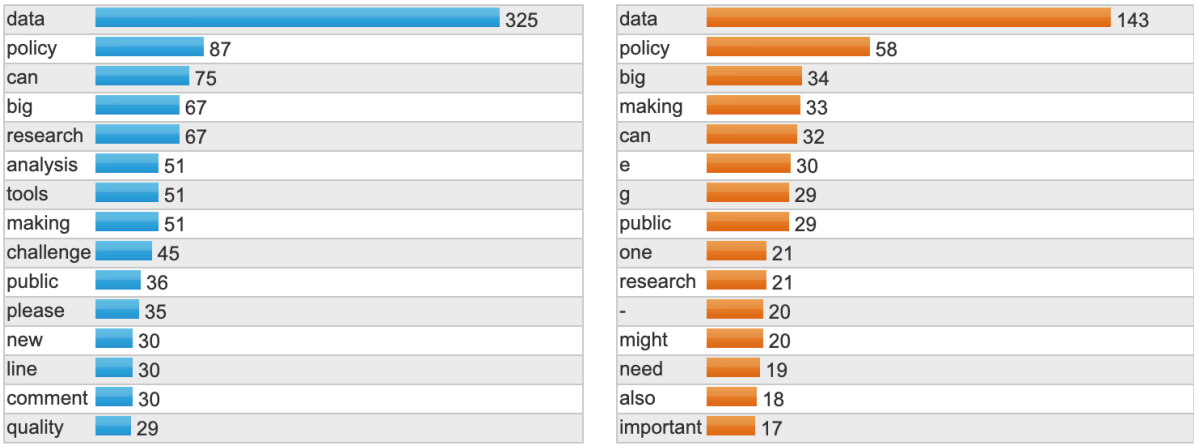

Figure 6 - Most used words in speech (left) and comments (right)

Below is also presented a table with examples of relevant comments and commenters.

Name and Surname	Position	Example of Comment
Alan Hartman	Senior Lecturer at Department of Information Systems, University of Haifa	Mediation between traditional objectives and boundaries with information from big data and participative democracy. Relevant for this problem are optimization methods, techniques like linear and non linear programming. Leonid Kantorovich and Paul Cockshoot (and the economic calculation debate) are a good reference for this topic, how to combine big data and economic planning. Cyber gis technology could be a further extension.
Basanta Thapa	Researcher at Fraunhofer FOKUS	Co-creation is not an inherent part of data-driven policymaking. It is perfectly possible (and widely practised) to conduct data-driven policymaking without involvement of other stakeholders. Please do not produce murky definitions of concepts.
Carlos Agostinho	Director of Operations at Knowledgebiz	This is a very interesting subject of discussion. Please refer to a very interesting publication: "Digital Transformation: Is Public Sector Following the Enterprise 2.0 Paradigm?"
Christos Botsikas	Information Technologist at National Technical University of Athens	Data coming from citizens can be of great, but also of zero value. Campaigns letting them know of the potential value of their data should be realised. And they should even be reimbursed (e.g. small-scale tax reliefs?) when they contribute with actually valuable data. Direct co-creation might not be the most effective and efficient approach.
Enrico Ferro	Head of Innovation Development Department at Links Foundation	You may consider using homomorphic encryption to train algorithms while preserving the privacy of the data owners.

Juliane Schmeling	Researcher at Fraunhofer FOKUS	Add the research cluster: "Public Governance frameworks for data driven policy making structures" The cluster can encompass different governance approaches: 1. approach: (ex post) continuous monitoring based on target- and indicator systems 2. approach: (ex ante) simulation of different political solutions
Luca Alessandro Remotti	Business Innovation Project Manager at Join Institute of Innovation Policy	The real issue, regularly highlighted, concerns the availability of data. One case is that the needed data is simply not collected and the level of granularity needed or targeted. Another case is that data is actually collected but since it is an asset by those who collect and detain it, it is not shared or shared at a (high) cost. Policy making typically would like to rely on open (and free) data.
Maria Wimmer	Professor at University of Koblenz-Landau	Where is the manipulation of statements and misinformation (distortion of meanings, etc.) in the www and identity theft threated?
Mariam El Ouirdi	Researcher at the University of Antwerp	Co-creation may not be inherent, but it may be increasingly inevitable. The European Commission talks about "the advent of co-responsibility with citizens and businesses (co-design, co-production, co-evaluation, etc.)". Reference: European Commission. (2017). Quality of Public Administration A Toolbox for Practitioners.
Yannis Charalabidis	Professor at University of Aegean, World's 100 Most Influential People in Digital Government	Allow for some not model-centric approaches in simulation. Machine - learning based instead of modelling based approaches. (where the machine uses neural nets and solves a problem, but no one really knows how ...)

A second version of the roadmap has been uploaded at the end of August and has collected 336 comments from 21 commenters. Below is also presented a table with examples of relevant comments and commenters.

Name and Surname	Position	Example of Comment
Gianluca Misuraca	Senior scientist at JRC Seville	An important aspect to considering is the dynamic development of the policy process and the need to allow for feedback loops. In other words it is important not consider policy making as a linear process, rather a complex - non-linear - adaptative cycle.
Akrivi Vivian Kiousi	Senior Business Development Manager, Head of Transport Lab Research and	Also a TT deriving recommendation to assist the development of new tools and techniques is to foster on the data quality at least for important datasets. Policy recommendations: Data Interoperability to foster collaboration

	Innovation at Intrasoft	
Angela Guarino	Policy Officer at the European Commission	Data visualisation and scenario visualisation should be tested in presence, maybe the "citizen juries" method can be used in order to make improvement and validate results.http://designresearchtechniques.com/casestudies/citizen-juries-an-action-research-method/
Anna Triantafillou	Deputy Head of Innovation Lab at Athens Technology Center	Social networks are full of bots, promoted posts etc. I don't know if they are the best source.
Evmorfia Biliri	Researcher at Fraunhofer FOKUS	Personal data are not that much of a barrier imo. I think that GDPR for example easily allows processing. Sensitive data is the key here.
Shefali Virkar	Research Associate at Donau-Universität Krems	See, for example: Rinnerbauer, B., Thurnay, L., Lampoltshammer, T. J. (2018). Limitations of Legal Warranty in Trade of Data. Virkar, S., Parycek, P. Edelmann, N., Glassey, O., Janssen, M., Scholl, H. J., Tambouris, E., Proceedings of the International Conference EGOV-CeDEM-ePart 2018, 3-5 September 2018. Danube University Krems, Austria: 143-151, Edition Donau-Universität Krems.
Spiros Mouzakitis	Researcher at National Technical University of Athens	IMHO, data-driven policy making is the key. The BIG data notion is a good-to-have. But first you need to be certain that the data-driven approach has been achieved.

Overall, the tool collected 667 comments, going beyond the KPI of 500 comments.

2.3.4 Input from other Events

The roadmap has been presented in the last Big Policy Canvas conference, as well as in four other events:

- Data for Policy 2019 – Digital Trust and Personal Data (London, 11-12 June 2019). The Big Policy Canvas roadmap was presented in the Session "Governance Technologies 1", chaired by Martijn Poel, from the Ministry of Education, Culture and Science, in the Netherlands. The other presentations in the session were:
    - "Imagining Futures – A generative scenario-based methodology to improve planning and decision-support systems for policymakers"; Vaibhav Dutt, Srijan Sil, Harsha Krishna and Bharath Palavalli - Fields of View, India;
    - "Trusted Smart Statistics: how new data will change official statistics"; Fabio Ricciato* and Albrecht Wirthmann – European Commission – EUROSTAT, Luxemburg;
    - "What is the role of a data-driven public sector in the well-being of citizens? Benjamin Welby* (@bmwelby) and Barbara-Chiara Ubaldi (@BarbaraUbaldi) - OECD, France (@OECD).

- EGOV2019 – Joint conference EGOV-CeDEM-EPART 2019 (San Benedetto del Tronto, 2-5 September 2019). The workshop presented and discussed the Big Policy Canvas Roadmap for Future Research Directions in Data-Driven Policy Making, which defines a set of 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. Francesco Mureddu is Chair of the Practitioners Track, while David Osimo is Chair of the Track Digital Society.
- "The 9th Samos 2019 Summit On ICT-enabled Governance" in conjunction with "The 6th International Summer School On Open and Collaborative Governance." (Samos, 1-5 July 2019). In the 7th Session: Workshop II on a Roadmap to Future Government. In this session we proceed to presentations and discussion concerning the development of the new roadmap for digital government. Organizers: Maria Wimmer, Koblentz University; Francesco Mureddu, Lisbon Council; Juliane Schmeling Fraunhofer Institut FOKUS; Shoumaya Ben Dhaou, United Nations University. It has to be noticed that the workshop has been co-organised with the consortium of the project Gov3.0, and that Francesco Mureddu gave classes based on the roadmap also in the Summer School on Open and Collaborative Governance.
- BDV PPP Summit – Impact Empowered by Data-Driven Artificial Intelligence (Riga, June 26-28 2019). The Big Policy Canvas roadmap was presented in a panel aimed to take stock of the current lessons learnt and the near future policy challenges for big data solutions. The panel was divided into three subthemes of each 30 minutes. These where: 'Big themes and big challenges', 'Data markets: lessons learnt and regulatory challenges' and 'Policy4Data and Data4Policy'. The roadmap was presented in the last session, and after the presentation there was a discussion and the attendants were invited to our website and to provide comments to our tool. The other panellists were: Marina Micheli (JRC), Theodora Varvarigou (National Technical University of Athens), Edwin-Morley-Fletcher (Lynkeus), Fernando Perales (JOT Internet Media), Vivian Akrivi Kiousi (Instrasoft), Karolina La Fors (E-Sides), Julien Debussche (Bird & Bird), and Mauricio Fadel (BigDataStack). It has to be noticed that the Big Policy Canvas team has contributed to a policy brief stemming from the event.

3 Current Status and What is New

3.1 The Policy Cycles

Policy-making is typically carried out through a set of activities described as "policy-cycle" (Howard 2005). In this document we propose a new way of implementing policies, by first assessing their impacts in a virtual environment. While different versions of the cycle are proposed in literature, in this context, we adopt a simple version articulated in 4 phases:

- **Agenda setting** encompasses the basic analysis on the nature and size of problems at stakes are addressed, including the causal relationships between the different factors;
- **Policy design** includes the development of the possible solutions, the analysis of the potential impact of these solutions, the development and revision of a policy proposal;
- **Implementation** is often considered the most challenging phase, as it needs to translate the policy objectives in concrete activities, that have to deal with the complexity of the real world. It includes ensuring a broader understanding, the change of behaviour and the active collaboration of all stakeholders. This phase includes also adoption, where accountability and representativeness are needed. It is also the area most covered by existing research on e-democracy;
- **Monitoring and evaluation** make use of implementation data to assess whether the policy is being implemented as planned, and is achieving the expected objectives.

Figure 7 below (authors' elaboration based on Howard 2005) illustrates the main phases of the policy cycle (in the internal circle) and the typical concrete activities (external circle) that accompany this cycle. In particular, the identified activities are based on the Impact Assessment Guidelines of the European Commission[10].

[10] Impact assessment guidelines SEC(2009) 92. Key documents are on the IA website (http://ec.europa.eu/governance/impact/key_docs/key_docs_en.htm).

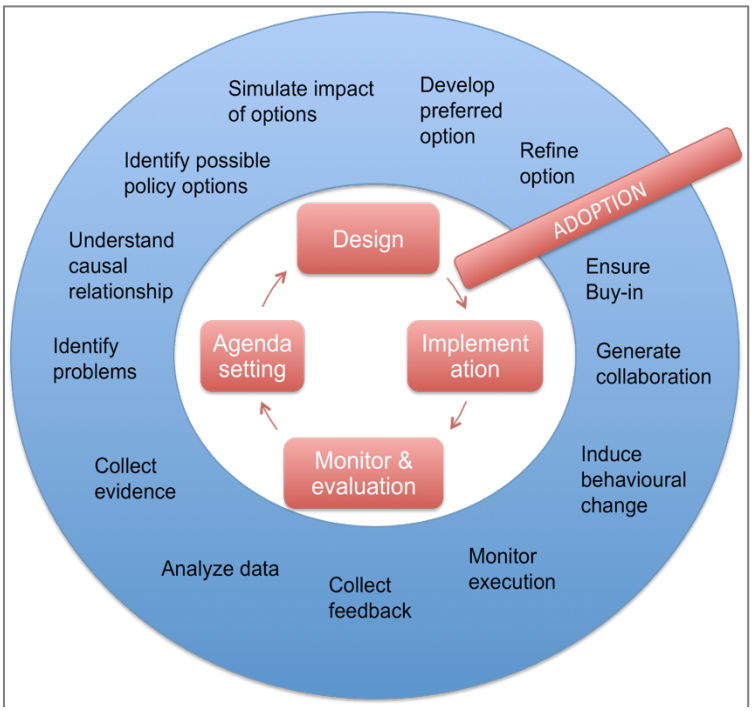

Figure 7 – Policy Cycle and Related Big Data Activities

This is clearly a mere approximation of the policy cycle framework, and more advanced versions exist, as for instance the e-policy cycle, which builds on the fact that by leveraging on Big Data Analytics, the evaluation can be carried out not only at the end of the policy cycle, but at any stage (Höchtl et al. 2016).

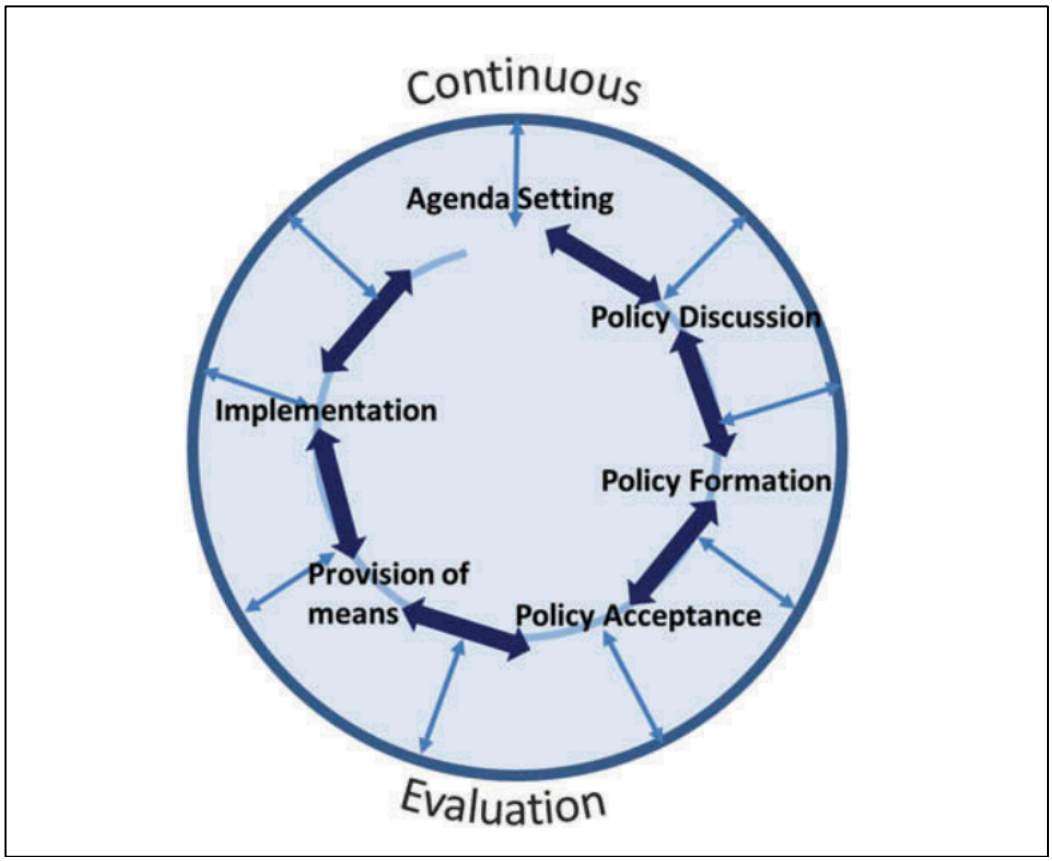

Figure 8 - The big data-revised policy cycle

Specifically in the evaluation phase, big data and data analytics approaches can help detecting the impact of policies at an early stage (), before formal evaluation exercises are carried out, or detecting problems related to implementation, such as corruption in public spending. Most importantly, big data can be used for continuous evaluation of policies, to inform the policy analysis process, while even empowering and engage citizens and stakeholders in the process (Schintler and Kulkarni 2014; Höchtl et al. 2016). Testing a new policy in real time can provide insights whether it has the desired effect or requires modification. Furthermore, as shown by Dunleavy (2016) big data can be used for behavioral insights. In this respect, the production of new data can also stem from the involvement of citizens science experiments, aimed at collecting data from the real world in real time. A thorough description of the various phases and the relative use of big data analytics is provided below.

3.2 The Traditional Tools of Policy Making

Let us present now what are the methodologies and tools already traditionally adopted in policy-making. Typically, in the agenda-setting phase, statistics are analysed by government and experts contracted by government in order to understand the problems at stake and the underlying causes of the problems. Survey and consultations, including online ones, are frequently used to assess the stakeholders' priorities, and typically analysed in-house. General-equilibrium models are used as an assessment framework. Once the problems and its causes are defined, the policy design phase is typically articulated through an ex-ante impact assessment approach. A limited set of policy options are formulated in house with the involvement of experts and stakeholders. For each option, models are simulated in order to

forecast possible sectoral and cross--sectoral impacts. These simulations are typically carried out by general-equilibrium models if the time frame is focused on short and medium term economic impacts of policy implementation. Based on the simulated impact, the best option is submitted for adoption. The adoption phase is typically carried out by the official authority, either legislative or executive (depending on the type of policy). In Some cases, decision is left to citizens through direct democracy, through a referendum or tools such as participatory budgeting; or to stakeholders through self-regulation. The Implementation phase typically is carried out directly by government, using incentives and coercion. It benefits from technology mainly in terms of monitoring and surveillance, in order to manage incentives and coercion, for example through the database used for social security or taxes revenues. The monitoring and evaluation phase is supported by mathematical simulation studies and analysis of government data, typically carried out in-house or by contractors. Moreover, as numbers aggregate the impacts of everything that happens, including policy, it is difficult to single out the impacts of one policy ex post. Final results are published in report format, and fed back to the agenda setting phase.

3.3 The Key Challenges of the Policy Makers

Let us now briefly discuss the key challenges which are faced by policy makers. One first aspect to consider is the emergence of a distributed governance model. Traditionally, the policy cycle is designed as a set of activities belonging to government, from the agenda setting to the delivery and evaluation. However, it has been increasingly recognized that public governance involves a wide range of stakeholders, who are increasingly involved not only in agenda-setting but in designing the policies, adopting them (through the increasing role of self-regulation), implementing them (through collaboration, voluntary action, corporate social responsibility), and evaluating them (such as in the case of civil society as watchdog of government).

Detect and understand problems before they become unsolvable

The continuous struggle for evidence-based policy-making can have some important and potentially negative implications in terms of the capacity of prompt identification of problems. Policy-makers have to balance the need for prompt reaction with the need for justified action, by distinguishing signal from noise. Delayed actions are often ineffective; at the same time, short-term evidence can lead to opposite effects. In any case, government have scarce resources and need to prioritize interventions on the most important problems. For instance, the significant underestimation of the risks of the housing bubble in the late 2000s, and the systemic reaction that it would lead to, led to delayed reactions. Systemic changes do not happen gradually, but become visible only when it is too late to intervene or the cost of intervening is too high. For example, ICT is today recognized as a key driver of productivity and growth, but evidence to prove this became available at a distance of 3-5 years from the initial investment. In fact, the initial lack of correlation between ICT investment and productivity growth was mostly due to incorrect measurement of ICT capital prices and quality. Subsequent methodologies found that computer hardware played an increasing role as a source of economic growth (see inter al. Colecchia and Schreyer 2002, Jorgenson and Stiroh 2000, Oliner and Sichel 2000). The problem is in this case is therefore twofold: to collect data more rapidly; and to analyse them with a wider variety of models that account for systemic, long term effects and that are able to detect and anticipate weak signals or unexpected wild cards.

Generate high involvement of citizens in policy-making

The involvement of citizens in policy-making remains too often associated with short-termism and populism. It is difficult to engage citizens in policy discussions in the first place: public policy issues are not generally appealing and interesting as citizens fail to understand the relevance of the issues and to see "what's in it for me". The decline in voters' turnout and the lack of trust in politicians reflects this. More importantly, there are innumerable cases where the "right" policies are not adopted because they are not politically acceptable. While the Internet has long promised an opportunity for widespread involvement, e-participation initiatives often struggle to generate participation. Participation is often limited to those that are already interested in politics, rather than involving those that are not. When participation occurs, online debates tend to focus on eye-catching issues and polarized positions, in part because of the limits of the technology available. It Is extremely difficult and time consuming to generate open, large scale and meaningful discussion.

Identify "good ideas" and innovative solutions to long-standing problems

Innovation in policy-making is a slow process. Because of the technical nature of issues at hand, the policy discussion is often limited to restricted circles. Innovative policies tend to be "imported" through "institutional isomorphism". Innovative ideas, from both civil servants and citizens, fail to surface to the top hierarchy and are often blocked for institutional resistance. Existing instruments for large-scale brainstorming remain limited in usage, and fail to surface the most innovative ideas. Crowdsourcing typically focus on the most "attractive" ideas, rather than the most insightful.

Reduce uncertainty on the possible impacts of policies

When policy options have been developed, simulations are carried out to anticipate the likely impact of policies. The option with the most positive impact is normally the one that is proposed for adoption. Most existing methodologies and tools for the simulation of policy impacts work decently with well known, linear phenomena. However, they are not effective in times of crisis and fast change, which unfortunately turn out to be exactly the situations where government intervention is most needed. This is especially true in case of economic crisis, as shown by the policies carried out to fight the financial crisis in 2008. But the need for new policy making tools is not limited to the economic realm: in the future it will become more and more important to anticipate non-linear potentially catastrophic impacts from phenomena such as: climate change (draught and global warming); threshold climate effects such as poles' sea-ice withdraw, out-gassing from melting permafrost, Indian monsoon, oceans acidification; social instability affecting economic well-being (social conflict, anarchy and mass people movements). The lack of understanding of systemic impact has driven to short term policies which failed in grasping long term, systemic consequences and side effects.

Ensure long-term thinking

In traditional economics, decisions are utility-maximising. Agents rationally evaluate the consequences of their actions, and take the decisions that maximize their utility. However, it is well known that this rationalistic view does not fully capture human nature. We tend to overestimate short-term impact and underestimate the long term. In policy-making, short-termism is a frequent issue. People are reluctant to accept short-term sacrifices for long-term benefits. Politicians have elections typically every 5 years, and often their decisions are taken to maximize the impact "before the elections". There is also the perception that laypeople are less sensitive to long term consequences, which are instead better understood by experts. Overall, long-term impact is less visible and easier to hide, due to lack of evidence and data. As a result, decisions are too often taken looking at short-term benefits, even though

they might bring long term problems. This is especially true in a period in which populists movements are taking control.

Encourage behavioural change and uptake

Once policies are adopted, a key challenge is to make sure that all stakeholders comply with regulations or follow the recommendations. It is well known how the greatest resistance to a policy is not active opposition, but lack of application. For instance, several programmes to reduce alcohol dependency problems in the UK failed as they excessively relied on positive and negative incentives such as prohibition and taxes, but did not take into account peer-pressure and social relationships. They failed to leverage "the power of networks" (Ormerod 2010). For instance, any policy related to reduction of alcohol consumption through prohibitions and taxes is designed to fail as long as it does not take into account social networks. In another classical example (Christakis and Fowler 2007), a large scale longitudinal study showed that the chances of a person becoming obese rose by 57 per cent if he or she had a friend who became obese. The identification of social networks and the role of peer pressure in changing behavior is not considered in traditional policy-making tools. In this regard, important work is carried out by the UK Behavioural Insights Team (BIT).[11]

Manage crisis and the "unknown unknown"

The job of policy-makers is increasingly one of crisis management. There is robust evidence that the world is increasingly interconnected, and unstable (also because of climate change). Crises are by definition sudden and unpredictable. Dealing with unpredictability is therefore a key requirement of policy-making, but the present capacity to deal with crises is designed for a world where crises are exceptional, rather than the rule. Each crisis seems to find our decision-makers unprepared and unable to deal with it promptly. As Taleb (2007) puts it, we live in the age of "Extremistan": a world of "tipping points" (Schelling 1969), "cascades" and "power laws" (Barabasi 2003) where extreme events are "the new normal".

Detect non-compliance and mis-spending through better transparency

In times of budget constraints, it is ever more important for governments to ensure that financial resources are well spent and policies are duly implemented. But monitoring is a cost in itself, and a certain margin of inefficiency in resources deployment is somehow understandable. Yet the cost of this mismanagement is staggering: for instance, in 2010, 7.7% of all Structural Funds money was spent in error or against EU rules. The EU Commission has managed to bring the error rate down, achieving 2.4% in 2017 (3.1% in 2016, 3.8% in 2015 and 4.4% in 2014). This means more than €97 of every €100 spent by the EU was free from error. But the European Court of Auditors considers a 2% error rate as the level below which errors are not regarded as having a significant effect. Thereby it would be crucially important to be able to avoid the mismanagement with anticipatory corrective actions.

Moving from Conversation to Action

The collaborative action of people is able to achieve seemingly unachievable goals: experiences such as ZooGalaxy and Wikipedia show that mass collaboration can help achieve disruptive innovation. Yet too often web-based collaboration is confined to complaints and discussions, rather than action. A typical example is the electoral debate, in which before the elections we see an explosion of activity in social

[11] See: https://www.bi.team/

media discussing about the different candidates and their possible programs. Unfortunately, most of the time such energy then fails to translate into concrete action in the aftermath of the elections.

Understand the impact of policies

Measuring the impact of policies remains a challenge. Ideally, policy-makers would like to have real-time clear evidence on the direct impact of their choice. Instead, the effects of a policy are often delayed in time; the ultimate impact is affected by a multitude of factors in addition to the policy. Timely and robust evaluation remains an unsolvable puzzle. This is particularly true for research and innovation policy, where the results from investment are naturally expected at 3-5 years of distance. As Kuhlmann and Meyer-Krahmer (1995) puts it, "the results of evaluations necessarily arrive too late to be incorporated into the policy-making process".

3.4 Big Data Driven Policy Making

Let us now discuss the use of Big Data in policy making, and in particular in the policy cycle. First, we are going to introduce Big Data, their market and value chain, then we are discussing the application of big data to the policy cycle.

3.4.1 Big Data Value Chain

"Data-driven innovation is a key driver of growth and jobs that can significantly boost European competitiveness in the global market.", was declared in the EC Strategy "Towards a common European data space"[12]. Not only is data produced, gathered and elaborated by an increasing set of stakeholders at growing rates both in the public and private sector, but a true knowledge economy can also only build on data understanding, data integration and data-driven predictions. The term big data is employed to stress the scale of the problem to be solved and is usually explained through the 4 V's model:

- Scale of data (Volume) - 2.5 quintillion bytes created every day;
- Streaming/real-time data (Velocity) - 18.9 billion network connections in 2016;
- Heterogeneous formats (Variety) - 30 billion pieces of content are shared on Facebook in 2015;
- Data uncertainty (Veracity) - poor data quality costs the US economy $3.1 trillion in 2016.

Furthermore, as argued by Klievink et al. (2016), building on several studies (Adrian, 2011; Chen et al., 2014; Davenport et al., 2012; Gantz and Reinsel, 2011; Hota et al., 2015; Janssen and Kuk, 2016; Mayer-Schönberger and Cukier, 2013; OpenTracker, 2013; Simon, 2013), there are five differentiating characteristics of big data:

- Use and combining of multiple, large datasets, from various sources, both external and internal to the organization;
- Use and combining of structured (traditional) and less structured or unstructured (non-traditional) data in analysis activities;
- Use of incoming data streams in real time or near real time;
- Development and application of advanced analytics and algorithms, distributed computing and/or advanced technology to handle very large and complex computing tasks;

[12]Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions "Towards a common European data space" (COM(2018) 232 final)

- Innovative use of existing datasets and/or data sources for new and radically different applications than the data were gathered for or spring from.

Addressing a big data application means taking into account one or more of those four dimensions, which in turn require distinct technologies and approaches. The open data meme, on the other hand, emerged to highlight the transparency and legal issues related to data access and sharing. According to the Open Definition by the Open Data Institute, open data is "information that is available for anyone to use, for any purpose, at no cost" or, in other words, "data that can be freely used, modified, and shared by anyone for any purpose". On the pure technological side stands the concept of linked data that, coming from the academic research of the Semantic Web community, refers to a set of techniques and practices to structure, interlink and publish data. This is enabled by leveraging Web technologies (e.g. HTTP URI identifiers and hyperlinks) and machine-readable formats (e.g. RDF, related languages and relevant data models such as Data Cube, DCAT and ADMS) in order to foster interoperability. Certainly, to represent different perspectives on data, any combination of big, open and linked data is possible: the expression "big open linked data" thus refers to large datasets suffering from one or more of the four V's issues, released with an open license for both commercial and noncommercial purposes, and published on the Web in machine-readable format and interlinked with other data sources.

Considering all these characteristics, as reported by AGCOM(217/17CONS) it is possible to highlight this radical change in the approach to data analysis. In fact, in the time of data scarcity it was necessary to ask a research question and consequently collect data ("data-is-scarce-model"), or to acquire a sample, looking for the answer to a predetermined research question. In the time of data abundance, data are often collected regardless of specific research questions, which are then defined a posteriori after the analysis (Figure 9)[13].

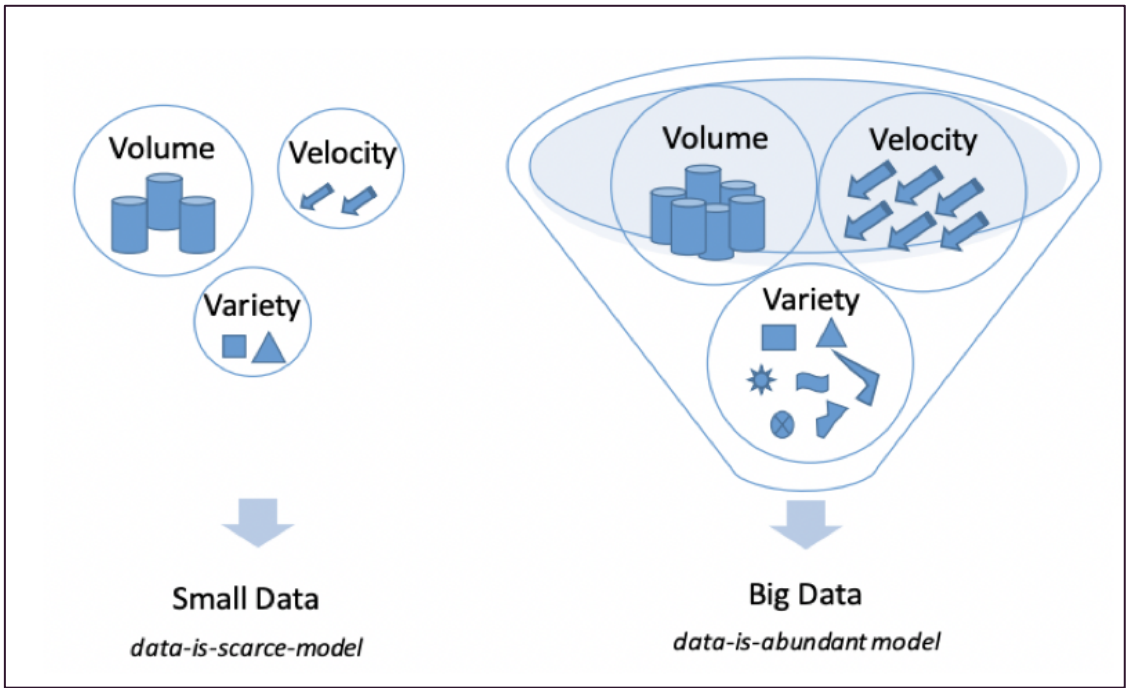

Figure 9 – Small and Big Data

The landscape of data analytics is very broad in terms of definitions, perspectives and actors. In the following, we give our definitions of the main topics – mainly in line with the definitions of the EU

[13] Source: AGCOM(217/17/CONS)

strategy – highlighting the current trends and technologies with special reference to evidence-based policy-making. By data analytics, we mean the set of approaches and methodologies to explore data with the purpose of drawing conclusions or taking decisions on top of that information. Data analytics as such is a complex process that encompasses collecting, organizing and processing of datasets, in order to discover hidden patterns and relations in data and to make predictions on future data. Data analytics usually aims at measuring business performance (KPIs), provide suggestions or support and inform decision-making. Thus, data analytics spans across different activities: from data analysis to data cleansing, from data processing to data modelling, and from data prediction to data visualization. To take the point of view of big industry players, IBM states that the purpose of analytics is to "discover what is happening, determine why it is happening, predict what is likely to happen and prescribe the best action to take". SAS interprets the current popularity of analytics technologies as a sign that "we are on the cusp of an analytics revolution that may well transform how organizations are managed, and also transform the economies and societies in which they operate". Ericsson has a vision of "Data-derived growth: creating innovative offerings and generating new revenue streams sparked by data analytics". Data analytics employs multiple types of data. Apart from the traditional distinction between structured data (e.g. databases) and unstructured information (e.g. text), different perspectives can be chosen to address data analytics.

Figure 10[14] illustrates a typical Big Data Value Chain and the respective technologies used in every step of the chain. While most of the techniques can be considered state-of-the-art (statistics, data mining, basic machine learning), the scientific-technical challenges – and the business opportunities – come from applying those and more sophisticated techniques to big data (in the sense of data fitting in the 4 V's model) and to contexts requiring data integration from multiple and heterogeneous sources (e.g. data that may require signal processing, NLP, spatiotemporal analysis and predictive modelling).

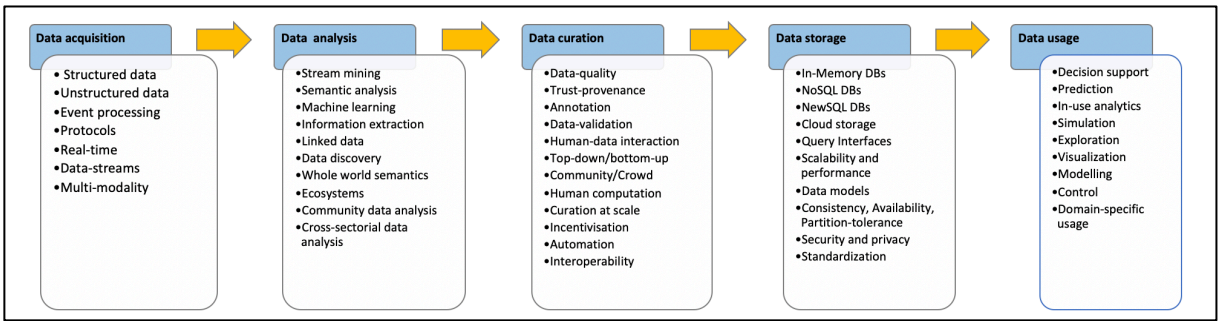

Figure 10 – Big Data Value Chain and Technologies

On the same line, in Figure 11[15], it is depicted the Big Data Ecosystem, which shows the interconnection among the following actors:

- Subjects generating data, i.e. data "providers";
- Technology providers, typically in the form of data management platforms;
- Users, i.e. who use the big data to create added value;
- Data brokers, which are organizations collecting data from a set of sources, both public and private, and that sell them to other organizations;
- Companies and research organizations, who develop new technologies, new algorithms through which to explore data and extract value;

[14] Source: Curry (2016)
[15] Source: AGCOM (217/17/CONS)

- Public bodies who act as regulators and/or public administration providing products and services to the citizens based on data, or that use data in their processes.

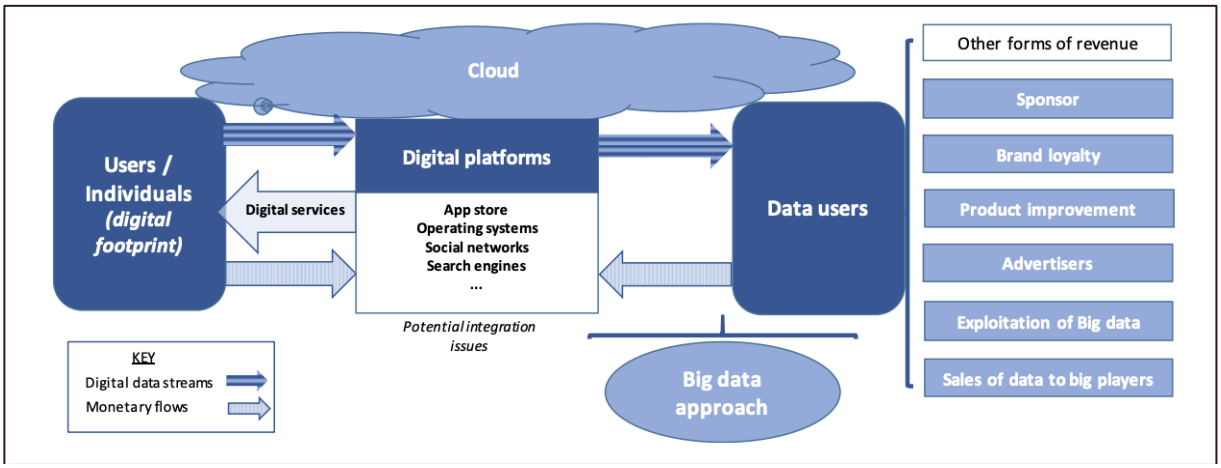

Figure 11 – Digital Data Market

The market of data analytics is continuously growing, even if marker research tends to focus on the big data aspect because it spans to a heterogeneous – and potentially expensive – set of solutions. Gartner affirms that 64% of organizations invested or planned to invest in big data in 2013[16]; 67% of the respondents to an MIT Sloan Management Review survey report that their companies are gaining a competitive edge from their use of analytics[17]; Forrester Research estimates that organizations effectively use less than 5% of their available data (Killmeyer et al. 2014). To take the point of view of a specific industry sector, McKinsey Global Institute calculates the potential value generated by big data solutions in US health care in $300 billion in 2010 (Manyika et al. 2011). Big data and data analytics offer unquestionable added value to a diverse set of sectors; hereafter we give some short examples. Multimodal transport and logistics need the integration of sensor data with vast amounts of mobility and social data generated by an increasing set of devices and technologies; the opportunities are monitoring, controlling and managing transport and logistics processes, for example with goods delivery adaptation based on predictive monitoring or weather forecasts. In the environmental sector, the need is to understand planet and climate changes, also considering the role of human factors and interventions; the challenge is to improve the accuracy and availability of location data (e.g. those of the European programmes Copernicus and Galileo) to the benefit of the industry at large. The media and content market is another field in which big data and data analytics can improve the dynamic access and interaction patterns of users with media, also in a "prosumer" (producer and consumer) approach, to reach a seamless content experience. In the energy and manufacturing industries, the digitization of processes and the increasing supply of sensor networks require new approaches and technologies to (real-time) data management which in turn can lead to efficiency gains, predictive maintenance and better resource consumption.

[16] More info available at https://techcrunch.com/2013/09/23/64-of-organizations-have-invested-in-or-plan-to-invest-in-big-data-tech-but-only-8-have-started-using-it-says-gartner/

[17] More info available at http://ilp.mit.edu/media/news_articles/smr/2015/56320Wx.pdf

3.4.2 Big Data in the Policy Cycle

The utilisation of Big Data technology and analytics for government is now in the early stages of a practical implementation (Moorthy et al. 2015). In fact, as reported by IDC[18], analytics alone will grow from $130.1 billion in 2016 to over $203 billion in 2020 among others driven by a shift towards a data-driven mindset. Furthermore, according to the Tech America Foundation[19], 82% of public IT officials say the effective use of real-time Big Data is the way of the future. While being still immature, big data technology solutions are applicable throughout the different phases of policy cycle, from agenda setting to policy design, implementation and evaluation.

The digital transformation of our societies, industry and government has made them unrecognisable in many ways. Many jobs have been lost, many more have been created. Costs of production of goods, provision of services, transport, and communication have significantly decreased, while speed, quality and efficiency have dramatically increased. However, the opportunities that digital technologies offer – notably with regards to public services – are yet to be fully seized. Indeed, an increased take-up of digital tools and solutions has the potential to render public services faster, cheaper, as well as more efficient, transparent, and user-oriented. The indirect positive effects are manifold, whether on public finances, on productivity, on citizens' lives, and on the environment. Overall, digital government transformation translates in a more competitive and attractive society. Today's society faces complex challenges such as migration, poverty, and climate change, for which not one optimal solution exists (Millard, 2015; Janssen and Helbig (2015). In order to address such problems, governments aim to realize public sector innovation that gears them towards becoming platforms of open governance, making optimal use of information and communication technologies (ICTs) to create public value (Millard, 2015). In this regard, the role of European governments has increased in complexity and complication. To cope with such challenges, ICTs have been increasingly used for enhancing the process of policy making and therefore address societal problems by formulating and implementing laws, rules and guidelines. In practical terms, data-driven policy making aims to make use of new data sources and new techniques for processing these data and to realize co-creation of policies, involving citizens and other relevant stakeholders. Clearly it is related to the notion of evidence-based policy making, which considers relevant the inclusion of systematic research, program management experience and political judgement in the policy making process (Head 2018). However, data-driven policy making stresses the importance of big data and open data sources into policy making as well as with co-creation of policy by involving citizens to increase legitimacy (Bijlsma et al 2011) and decrease citizens' distrust in government (Davies 2017). In this regard, a reported by Höchtl et al. (2016), data analytics has significant potential to be used in the policy cycle by contributing to policy decision making, in particular for what concerns:

- Identifying underperforming areas of public services and help with reallocation of resources for optimisation of public service provision;
- Improving existing processes by providing solutions for the citizens faster and with less paperwork;
- Predictions and forecasts.

[18] Please refer to https://www.ciodive.com/news/big-data-business-analytics-revenues-to-hit-203b-in-2020-idc-says/427507/

[19] TechAmerica Foundation. Demystifying Big Data: A practical guide to transforming the business of government. Retrieved from http://www.techamerica.org/Docs/fileManager.cfm?f=techamerica-bigdatareport-final.pdf

To achieve these benefits, it is however necessary to address the privacy and security issues arising from government handling vast amounts of citizen-related data (Höchtl et al., 2016).

Policy making is the process of creating and monitoring policies to solve societal challenges. In this respect, it is often conceptualized as a policy cycle, consisting of several different phases, such as agenda setting, policy formulation, decision- making, implementation and evaluation. Concerning the use of big data technologies in the policy cycle, according to Maciejewski (2017), big data supports better policy development and execution "by strengthening the information input for evidence-based decision-making and provides more immediate feedback on policy and its impacts". According to Schintler and Kulkarni (2014), big data has great potential as a resource for helping to inform different points in the policy analysis process "from problem conceptualization to ongoing evaluation of existing policies, and even empowering and engaging citizens and stakeholders in the process".

In this respect, the phases identified by Höchtl et al. (2016) are the following:

- **Agenda setting**: here, the challenge addressed is to detect (or even predict) problems before they become too costly to face[20], as well as reaching an agreement of which issues have to be dealt with. In this regard, through data governments can identify emergent topics early and to create relevant agenda points collecting data from social networks with high degrees of participation and identifying citizens' policy preferences. An important role is also played by the media, which are able to frame issues and spread relevant information (McCombs and Shaw 1972; Scheufele 1999);
- **Policy discussion**: this deals with debating the different options on the table, and identifying which is the most important. In this regard, opinion mining and sentiment analysis can help to inform policymakers about the current trend of the political discussion as well as the changes in public opinion in the light of discussed and proposed changes (Alfaro et al. 2013);
- **Policy formation and acceptance**: big data and data analytics solutions can be used for providing evidence for the ex ante impact assessment of policy options, by helping to predict possible outcomes of the different options, by making use of by advanced predictive analytics methodologies and scenario techniques. In this regard, Giest (2017) argues that the increased use of big data is shaping policy instruments, as "The vast amount of administrative data collected at various governmental levels and in different domains, such as tax systems, social programs, health records and the like, can— with their digitization— be used for decision-making in areas of education, economics, health and social policy". Another example would be the use of big data analytics to analyze and prevent the spread of disease (Harris 2015). Robust and transparent predictive modelling and algorithmic techniques can also help in improving the policy acceptance;
- **Provision of means**: here, the challenge is to improve the decisions on how to most efficiently provide the required personnel and financial means for the implementation of new policies by analyzing in detail past experiences. An example is given by use of big data in budgeting to increase efficiency and effectiveness while reducing costs (Manyika et al. 2011);
- **Implementation**: big data and data analytics can help identifying the key stakeholders to involve in policy or the key areas to be targeted by policies. Another way in which big data can influence the implementation stage of the policy process is the real-time production of data. The

[20] For instance, according to Longo et al. (2017), big data can serve as an input for "framing a policy problem before it is apprehended as such, indicating where a need is being unmet or where an emerging problem might be countered early" (p. 83).

execution of new policies immediately produces new data, which can be used to evaluate the effectiveness of policies and improving the future implementation processes.

In summary, big data tools and technologies present interesting opportunities to address some of the aforementioned key challenges of data-based policy making:

- Anticipate detection of problems before they become intractable;
- Generate a fruitful involvement of citizens in the policy making activity;
- Making sense of thousand opinions from citizens;
- Uncover causal relationships behind policy problems;
- Identify cheaper and real-time proxies for official statistics;
- Identify key stakeholders to be involved in or target by specific policies;
- Anticipate or monitor in real time the impact of policies.

However, data-driven policy making raises also complex challenges related to the capturing, integration and reuse of data exist (Bertot and Choi 2013, Janssen et al. 2012), as well as to the involvement of citizens and other stakeholders in policy making (Janssen and Helbig 2015, Ferro et al. 2013, Linders 2012). Furthermore, the assumption that simply because of the emergence of new technologies bureaucracies and public administrations will quickly adapt does not necessarily lead to the expected transformational outcomes. Big data readiness is an important factor, in order to avoid breaches of privacy and security of personal data, unfair treatment of citizens through overly extensive and unethical datafication of decision-making processes, wrong or suboptimal decisions because of incorrect data handling, analyses and interpretation (Clarke 2016, Janssen and Van den Hoven 2015, Margetts and Sutcliffe 2013).

Summarizing, several important research and implementation questions still need to be addressed:

- What is the disruptive and transformation potential of big data technologies for public sector operations and policy making activities?
- Which other activities do we have to consider - upskilling of personnel, changes, and adaptation of potentially outdated regulations, and investments in infrastructure?
- What are potential pathways and roadmaps that are to be followed by European public administrations consider starting the transformative processes of their own policy making activity?
- At the same time, how can we ensure that privacy, ethical and legal considerations are not jeopardized by these new technologies?

3.4.3 Bottlenecks and Enablers of Data-Driven Policy Making

It is widely accepted that data-driven policy making leads to better, more impactful policies. Governments need to be rigorous and responsive, thus having a solid evidence base is essential for an effective an improved policy analysis, that leads to consequent improvements in the service delivery and problem-solving capacities of the public administration. However, the systematic use of data to help public authorities in this process remains relatively rare, mainly because of lack of access to high quality data, lack of professional development in using data and a lack of collaboration around the use of data. Understanding the technical, organizational, social and business enablers and inhibitors that contribute to the effective use of big data is key to providing organisations with guidance for increased success in big data projects. In this respect, we have identified several external factors that can influence the adoption of big data technologies for the implementation of data-driven policy making:

Technology allows the collection and processing of huge amounts of data. The quality of data itself plays a key role in the implementation of big data strategies in the public sector. New trends like digital and mobile services, smart cities and the internet of things, together with the availability of new techniques and technologies that leverage today's available computing power to process vast amounts and varieties of data (social data, machine data and transactional data) in almost real time, are enabling evidence-based policy making practices. They offer a chance to be more citizen-focused, to include people's needs, actual behaviour, preferences and sentiment and satisfaction, as recorded for example on social media platforms. Open data, big data and data analytics are an opportunity to find insights in new and emerging types of data and content, as they represent a change in the quality, quantity and type of data public administrations handle. However, as the volume of data is growing exponentially in recent times, the risks of data deluge, inaccuracy or poor-quality data and the impact of incorrect data analyses increase as well. Data quality problems range from data that is not formatted properly without unique identifiers, data duplicates, missing data and misclassification to a poor control of data quality at the entry point. Besides, the amount of information collected about citizens can lead to privacy concerns and violation of the data protection regulation, apart from the danger of discrimination and stigmatization of certain part of the population. There is also the risk of error in the models used for big data methods, that may result in high volumes of wrongful actions, leading to high operational costs and high social costs for the people and businesses subject to such operations.

But public administration often has to deal with low budgets and legacy systems that do not leverage the power of such amount of data. Legacy IT, inflexible on-premise systems that are difficult and costly to maintain and improve, is still prevalent in the public sector, mainly because a lack of resources and the reliance on a technology that has been working for decades. Nevertheless, to ensure effective use of big data, system capabilities must meet the characteristics of big data, that exceed a normal system's capabilities in terms of scalability and performance.

It is not only about isolated data, but alto about being able to interoperate with different sources of data

In the EIF and the European Interoperability Strategy (EIS)[21] adopted by the European Commission in 2010 and coordinated through the ISA² programme, four layers of interoperability are identified where public administrations may face some challenges in the implementation of big data policy making activities:

- Legal interoperability, means the alignment of legislation allowing data to be exchanged according to commonly recognised rules and with a commonly agreed legal weight. There is generally a lack of common rules on data privacy and security requirements. The aggregation of data across administrative boundaries on a non-request-based manner (M2M communication) entails a real challenge, since this information may reveal highly sensitive personal and security information when combined with various other data sources, not only compromising individual privacy but also civil security. With the GDPR, citizens have more possibilities to better manage their personal data. It is especially important for governmental organizations to be transparent about the use of personal data and to provide citizens with a choice to make their personal data available;
- Organisational interoperability, defined as alignment of organisations and processes allowing to achieve the common goals ofthe cooperating organisation. Public sector organisations usually work differently with data or have in place different policies and principles. According to the Report on

[21] https://ec.europa.eu/isa2/eif_en

Organisational Interoperability and Public Service Governance[22], there is a need to build a common vision and goals across different levels of the public administration;

- Semantic interoperability, concerns the precise meaning of exchanged information as well as common definitions which are preerved and understood by all parties. The size and wealth of various big data sources implies a big challenge to make sure data is comparable. There is a need for documentation, overview and definitions to define the meaning of different terms in a specific context so they can be reused by different departments;
- Technical interoperability, concerns the alignment on technical elements involved in linking systems and services allowing data to be safely exchanged. In the public sector, data is usually fragmented, spread across not integrated systems and belonging to various departments. These data silos cause the problem of the timely access to data assets, because the data is not physically located where it needs to be used, processed, edited or analysed. Data silos tend to arise naturally in large organizations like public administrations because each organizational unit has different goals, priorities and responsibilities. Data silos can also occur when departments compete with each other instead of working towards a common business goal. When ISA² ends, additional funding activities of interoperability will come from the new programme Digital Europe, one of the committee's proposals in the multiannual financial framework for 2021-202723[24].

There is a lack of a clear strategy and leadership to incorporate big data insights into the policy making process. It takes time, effort and a change in the mindset to get systems, processes, staff capacity, and partnerships in place to successfully capture insights from administrative data. Policy managers are most of the times motivated by perceptions about external support, worried about interplay of diverse stakeholder values and interests. Political leaders are often preoccupied with maintaining support among allies, responding to press comments and social networks, polishing leadership credentials, and continuing established practices. In this context, even in the cases where data is collected, different kind of databases with different types of data are often linked or merged, large amounts of data are simply stored, and it may only become clear what the value or potential use of that data is after it has been collected. This may result in incorrect assumptions about the data, leading to poor quality insights. Cultural change is usually the most critical element in making an organization more data-driven. There is a lack of leaders that generate consistent enthusiasm across the wider organization. This enthusiasm needs to be embedded in the way people think and work every day, using data to guide any process and quality improvement.

And a shortage on European data scientists. According to the European Commission[25], the number of data workers in Europe will increase up to 10.43 million, with a compound average growth rate of 14.1% by 2020. The EU forecasted to face a data skills gap corresponding to 769,000 unfilled positions by 2020 and suggests that 100,000 new data-related jobs will be created in Europe by 2020. To support the policy making process, data worker profiles are demanded to convert data into useful information about causal patterns, trends over time, and understanding the likely effects of various policy instruments. Not only technical staff able to handle, analyse and report on big data sets is needed, also business-oriented profiles with the ability to ask the right questions, synthesize and leverage new data points quickly. In

[22] https://ec.europa.eu/isa2/news/report-organisational-interoperability-and-public-service-governance-published-participate-our_en
[23] https://ec.europa.eu/commission/sites/beta-political/files/budget-june2018-digital-transformation_en.pdf
[24] https://ec.europa.eu/commission/sites/beta-political/files/budget-june2018-digital-transformation_en.pdf
[25] https://ec.europa.eu/digital-single-market/en/news/final-results-european-data-market-study-measuring-size-and-trends-eu-data-economy

the public sector, the lack of digital literacy among government executives, as well as less competitive salaries, are the root cause of some of the challenges public-sector organizations face to succeed in the digital era, which raises serious challenges concerning the recruitment, training and retention of skilled analytical staff.

Having all this context in mind, we can identify several enablers for an effective use of big data in the public sector, while recognising future prospects and designing directions. We have grouped them in three different headings, attending to the external factors identified, although all of them are closely intertwined.

People and Organizational, that deals with the top management clear strategy and support, a culture of collaboration and having the right skills. The adoption of a data-driven policy making approach in a public administration should follow a strategy that should be well communicated to internal and external stakeholders, as well as to the citizenship in general. Successfully messaging the importance of data and evidence to achieve benefits can be a powerful tool building support for a new policy agenda, providing a path for various stakeholders to work together towards the common goals. For example, highlighting public administration commitment to efficiency by delivering quality government processes at the lowest cost and the social impact that can be achieved, promoting initiatives for offering new or better services to the citizens, etc. Leadership and board sponsorship of data programs is crucial, so the data-driven strategy should be backed with an adequate budget. Attractive business cases for big data analytics should be created and the related cost-benefit analysis needs to be performed to measure the potential long-term benefits of adopting a culture of making data easily available and support data-driven decision making. Findings should be also well communicated to non-technical stakeholders. It is important to get the adhesion of everyone in the business to adapt to new ways of thinking and working based on what the data is telling. For data to be transformational, it should not be restricted to managers. On the contrary, it needs to be business-critical, from the top to the bottom of the organization. The goal should be to get good data out to be readily accessible, interpretable, and actionable at the front line. And to truly embed data into mainstream thinking and behaviour, organizations need to look at how they can build data insights into business processes by default. Therefore, there should be sufficient well-educated data scientists who operate these tools and who are able to interpret the results in the correct context. If this requirement is not met, the potential of data cannot be fully seized. Some organisations may need to develop their existing talent, bringing people up to speed with the latest data trends and insight-led decisions, but a balance between 'in-house' skills and reliance on external advice from policy-oriented consultancy firms, local universities and independent think-tanks outside the public sector can foster a good culture of collaboration. The goal is to encourage people with different functional expertise coming together to complement their skills and work towards a common goal. Governments can even look to the private sector to find useful methods for building evidence to apply to their own policy making process, embedding the monitoring and testing of outcomes and metrics into decision making and service delivery processes and use that information to determine program effects and improve upon them. To successfully implement big data analytics initiatives, public sector organisations have to integrate data analytics in multiple processes, providing insight and adding value throughout various steps of the policy process and to multiple stakeholders. Besides, to realize this value from big data, governments must strengthen technical and legal frameworks to access and use data responsibly preventing inappropriate use of the big data by people with legal access to them. Public sector managers need to develop integrated capabilities to put big data insights into action, and to be

responsive to citizen feedback on services and polices to deliver societal value with big data analytics. Although public authorities do not relish being exposed to strong public criticism when program outcomes are disappointing or when pilot schemes produce very weak results, building a culture of evaluation is crucial, so political and organisational incentives for utilising evidence-based analysis and advice should be pursued.

Having the right data, meaning understand the purpose of collecting that data, ensure data quality, privacy and security. Big data has enabled countless of opportunities for those able to utilise it and can be used for many beneficial purposes by governmental organisations. On the other hand, the same technology could also be used to increase inequality and threaten democracy, therefore, the attention from the research community and policy makers should be placed on policies to prevent the possible misuses of technology. Public sector authorities must gain the confidence of citizens in the way big data is applied to the policy cycle. The pillars on which the public trust should be built are the quality and accuracy of the data, the security and protection of the data and the integrity of their use. Only some of these aspects are regulated in the European Data Protection Regulation (GDPR), the collection (with authorisation), management (with protection) and use (with limited authorisation and with licit purpose). Ethics must be implicit in the entire life cycle of the data, beginning with its collection and following with the most complex algorithms used for analyses and interpretation (AI). It is a mistake to think that, in comparison with human subjectivity, data is always accurate and objective. Data can be biased, poorly combined, manipulated or simply misinterpreted because it was based on erroneous algorithms. In consequence, implementation of big data processes into the policy making cycle must always be accompanied by sufficient guarantees that give the prohibition of discrimination, transparency and the reliability of both the data, as well as the analytical methods, the utmost relevance. This way, sufficient public trust can be achieved among citizens in the way the government uses big data possibilities in the policy cycle. It is also necessary to address the privacy and security issues arising from government handling vast amounts of citizen-related data and data integration practices. Public administrations tend to be protective of administrative data, often citing privacy concerns and institutional risk. Privacy concerns are paramount and should be deeply considered, however, public administration staff can easily misunderstand or too strictly interpret privacy laws. Increasing staff knowledge of and comfort with privacy laws can help to understand that data and information management, and the creation of qualitative meta-data on their public data assets like quality, structure, precise definition of content and insights obtained, is no longer only important for their own organisation. It might have to be exchanged with external parties and other governments who will use it to enrich their analysis and in overall provide better services. To deal with technical and legal challenges, the creation of common secured technology environments will be relevant to exchange big data files and implement a proper governance framework to address security and privacy concerns.

Supporting infrastructure and processes in place. Public sector has to deal with budget pressures and do more with less. Although it can be seen as a short-term expenditure, modernise legacy systems is a powerful way to improve efficiency, as siloed legacy systems, that may not support linking data points across common fields, can hinder innovation by not only restricting access to core data or limiting information sharing across departments and organizations, but also by consuming budget and resources for maintenance. Big data, due to its own nature of volume and variety, is constantly growing, so it is relevant to invest in a solution architecture that can scale on various components and that provides good performance and user experience. Besides, to select the right software tools and architectural

components, all specialists' requirements must be taken into account. It is necessary to convince budgets managers of the need to invest in technology, looking for new arguments that allow to show the value of big data with facts and evidences. Experimentation environments, where technology can be tested and data visualized, can be useful to provide the justification of technology investments based on business value. After a successful experiment, organisations might be willing to incorporate them in their regular operations and processes. To split the upfront investments, collaboration models can be advisable to get extra benefits from sharing learning experiences, economy of scale and the economic and social value creation of data outside the public sector thanks to the potential reuse of open data. In conclusion, in the public administration big data can be used as input to processes aimed at gaining new insights to enable better ways of policy making. Governments generate and collect vast quantities of data through their everyday activities, citizens give their data away through new forms of participation with their smart phones, contained in videos, images, or textual information exchanged in social networks, there are billions of devices that can sense, communicate, compute and potentially actuate. Data on inputs, outputs, productivity and processes can all be captured and recalled in more comprehensive detail than ever before but even with the best data, policy making is not an exact science, thoughtful human analysis is required to interpret available data and incorporate factors that may not be reflected in the data collected. This is especially relevant in the public sector aiming at enhancing the policy making processes.

4 Identification of Gaps and Research Needs

4.1 Step 1: Needs Selection

In the Knowledge Base[26] there are 28 Needs identified, from those, 12 are of High Priority, and 11 of them have big data potential. After we apply the selection of Needs with High Priority and big data potential from the Knowledge Base filters we end-up with the Needs listed in Table 3.

Table 3 - Needs selected for Gap analysis

Reference	Priority	Type	Name
N-S-1	High	Strategical	Development of domain specific target and indicator systems
N-S-2	High	Strategical	Involvement of the public and citizens, as well as the development of citizen-centered policy making
N-S-4	High	Strategical	Strengthen citizens' trust in public administration
N-S-9	High	Strategical	Cross-linked information exchange
N-O-7	High	Organisational	Standardisation of processes
N-T-1	High	Technical	Cope with the production of huge volumes of data
N-T-3	High	Technical	Ensuring data security taking into account the protection of citizens' privacy
N-T-4	High	Technical	Establishment of a comprehensive technical infrastructure and IT architecture
N-I-1	High	Informational	Link between impact, quality, performance measurements and financial information
N-I-3	High	Informational	Ensure availability of (real-time) information and knowledge
N-I-4	High	Informational	Comprehensive knowledge and information management

From the eleven Needs, four are of strategical type, one organisational, three technical and three informational.

4.2 Step 1 to Step 4: Need Breakdown, Asset Assessment and Gap Identification

This set of steps is performed for each Need individually. Each Need is broken down into its functionalities, the assets are assessed against those functionalities, and the gap is identified for each functionality. For each Need, a table is shown with the text of the Need, the fist *functionality* and its corresponding *gap* below, and below a second set of *functionality-gap* in case there is one. Tables with the assessment of the assets against the functionalities are available in **Error! Reference source not found.**.

Table 4 - Gap identification for N-S-1, Development of domain specific target and indicator systems

N-S-1	Development of domain specific target and indicator systems

[26] Big Policy Canvas, Knowledge Base, https://www.bigpolicycanvas.eu/community/kb, retrieved March 2019

Already the political economist and sociologist Max Weber once has pointed out that decision makers need to ensure the rationality of their decisions, by trying to balance out the best relation of means and ends (Weber 1980). Consequently, policy makers need to clarify the targets that they want to reach through certain political programmes and norms. In fact, the executive bodies need quite precise targets, since they are responsible for the adoption and implementation of political and legal solutions and need to translate political solutions in concrete activities. If public administrations want to monitor political targets, they need to set up a management control system, as it is already quite common in the private sector. Nevertheless, since it is not possible to score success from insulated financial ratios (See also Need: Link between impact, quality, performance measurements and financial information), the public sector needs to observe much more complex systems in consideration of public interests (Budäus and Buchholtz 1997). In a conducted interview with a public administration representative on the regional ministerial level in the youth welfare policy domain, the interviewee confirmed that there is a lack of clearly formulated goals on the political level. The interviewee further mentioned that without clear goals on a political level, executive bodies are incapable to derive operationalised goals and indicators. A second problem he mentioned is that targets, if they are formulated, should be well balanced among each other, since it is important in the implementation phase to know which targets have priority to set up a strategic planning. For example, it is difficult to implement child day care availability for everybody and best trained childcare workers at the same time. To sum up, policy domain specific targets and indicator systems are especially relevant in the formulation, and implementation phase, but are also relevant in the monitoring and evaluation phase, since it is impossible to monitor and evaluate political targets and their derived indicators in a performance measurement system without targets.	
Functionality N-S-1.F1	Management control system to monitor political targets based on multiple indicators (impact, quality, performance measurements and financial information).
Gap N-S-1.F1.G1	Lack of an approach to develop socio-technical control systems tailored to specific domains and its specific KPIs. So far, tailor-made systems or assessment frameworks have been developed. A methodology to build such standardized and modular systems is required.
Functionality N-S-1.F2	Definition of clear goals in policy building with balanced targets.
Gap N-S-1.F2.G2	Each domain requires proven approaches to define the goals and balanced targets to monitor the application of policies.

Table 5 - Gap identification for N-S-2, Involvement of the public and citizens, as well as the development of citizen-centred policy making

N-S-2	**Involvement of the public and citizens, as well as the development of citizen-centred policy making**
Concerning the public, a close cooperation between public administration and citizens seems essential. Through participative democracy and public involvement, a new relationship between the citizens and the administrations can be established. The publicity becomes a valued partner to identify problems, discover new thinking and propose solutions. This can be seen as a profit for public administrations, because the experiences of the citizens can be contributed into the administration and help to improve, for example, its policy making. As the main customer of public administration, the wishes and needs of the citizens (customer satisfaction) should be involved in policy making and be automatically transferred into administrational needs. It is of big interest what the customer is thinking and what the customer wants. This can lead to improvement of efficiency and effectiveness. The changed living environment of the customers (internet, online shopping, 24-hour availability of products) raises the expectations towards the administration, which must meet these demands. Engaging the public can help to rebuild the trust of citizen and consequently lead to a stronger citizens' satisfaction (Thomas 2013).	

Functionality N-S-2.F1	Participative democracy.
Gap N-S-2.F1.G1	Realistic participation is very limited, mainly through games and simulations or comment and suggestion tools. There are some tools that provide support to participative democracy but there is a lack of real experiences where direct democracy is applied, and when found it is very limited in its scope. Conclusion, there are no experiences that really exploit the full potential of participative democracy.
Functionality N-S-2.F2	Improvement of efficiency and effectiveness by transferring to PAs the experiences, wishes and needs of the citizens into administrational needs in the policy making process.
Gap N-S-2.F2.G2	Tools and applications are mainly devoted to idea collection and scoring or problem and idea reporting mainly for environment in urban areas. So specific transfer of needs (all kind) to PAs is lacking. It is not a problem of tools availability, but of political willingness to really involve citizens in development of citizen-centred policy making.

Table 6 - Gap identification for N-S-4, Strengthen citizens' trust in public administration

N-S-4	**Strengthen citizens' trust in public administration**
To improve public administration´s image, it is important to rebuild the trust in it. The citizens' cooperation seems essential to achieve public purposes. The lack of trust can make the formulation and implementation of policies more difficult or even impossible. Relevant factors that influence citizens' trust is the administrations´ integrity, as well as its performance. Transparency and public participation can be helpful possibilities to increase the trust in government and administration (Grimmelikhuijsen et al. 2013, Wang and Van Wart 2007, Olabe 2017). The need has not been validated in the qualitative interviews, but seems to have relevance for the public sector due to the findings of the desk research. This need is a key in the policy formulation phase, because only with the trust of the population, problems can be understood consequently right and the necessary policies can be developed.	
Functionality N-S-4.F1	Citizens' cooperation.
Gap N-S-4.F1.G1	The citizens' acceptance of public participation offers often falls short of expectations because participation requires time and other resources (technical, logistic, educational). There are educational games at the citizens disposal but not sufficiently promoted or not accessible for all audiences.
Gap N-S-4.F1.G2	Appropriate legislation at the Member States level which can better integrate public consultation, as policy-making is more based on expert inputs in detriment of non-expert knowledge coming from other parts of society. Therefore, there is a need to change institutional and organizational culture and a shift of mindset from civil servants, policy makers, politicians, researchers but also from citizens and civil society.
Functionality N-S-4.F2	Transparency
Gap N-S-4.F2.G3	Citizens tend to refrain from engaging in participation procedures because it is not clear for them what will happen with their contributions and because trust in political bodies is lacking (sometimes it is not always the interests of the most concerned that become accepted but often the interests of the best organized). The issue of representativeness needs to be preserved through a mix of tools and methods that ensures a good variety of viewpoints. This can be reinforced by more transparency about the use and influence of citizens' feedback, thus avoiding concerns about potential conflict of interests or biased collection of inputs.

Table 7 - Gap identification for N-S-9, Cross-linked information exchange

N-S-9	Cross-linked information exchange
Public sector organisations are mainly knowledge-intensive organisations, and to exploit their knowledge, effective knowledge sharing among the different departments is required. There can be great advantages if information is not only used in the own administration but is shared between hierarchies, different policy areas and levels of government. Including findings from other disciplines in respective monitoring systems (e.g. education, social, youth, and work) can create synergy and learning effects, which in turn leads to a share of benefits. In the interview with the division head in the policy domain "Youth and Welfare" on regional level, it became clear, that the information exchange is a big issue in German administrations. Due to the federal structure, the data belongs to different players and cannot be easily matched. Analyses and comparisons are made more difficult, whereby valuable information is lost. Especially in the agenda setting and implementation phase, cross-linked information exchange can bring valuable improvements.	
Functionality N-S-9.F1	Share information among hierarchies, different policy areas and levels of government, creating synergies.
Gap N-S-9.F1.G1	While there are a lot of initiatives to share public open data for citizens / businesses, there is a gap in sharing information internally (also closed data) between different public administrations. Interoperability is often lacking and most of the times political will. Also, ignorance about the availability and benefits of sharing data in the administration. The publication of open data is usually considered more a duty (additional workload perception) that an opportunity, and possible reuse is more focused to private companies that own administration.
Gap N-S-9.F1.G2	There is a lack of alignment with the interests of re-users, due to a lack of communication between suppliers and users. The data published has not always the necessary quality, it is not updated frequently or there is no homogeneity between different countries, and this makes reuse inefficient and costly.
Gap N-S-9.F1.G3	Intellectual property rights and the diversity of licences makes information sharing and reuse more difficult.
Gap N-S-9.F1.G4	Sharing information can lead to an eventual improvement, but there is a huge trade-off between the opportunity to improve and to reveal that processes are implemented in a non-optimal manner (fear of criticism). Sharing of process knowledge has the potential to greatly improve the organization but is connected with a high degree of self-exposure and risk.
Gap N-S-9.F1.G5	Lack of altruistic culture and conservative pattern of behaviour in public administration.

Table 8 - Gap identification for N-O-7, Standardisation of processes

N-O-7	Standardisation of processes
Standards require a certain legal basis and binding specifications. At the same time, they must be also accepted by the target group. If standards are enforced, they offer the advantage of planning and investment protection. This provides a good basis for further digitisation of processes (Groppo and Heck 2009). As an example, interviewed employees of the administration referred to a nationwide same process for which there are different procedures in all municipalities. In addition, the media interruption between the administration and external partners was criticised. Standards can optimise these processes, increase the efficiency and save time.	
Functionality N-O-7.F1	Legal basis & specifications.

Gap N-O-7.F1.G1	Initiatives are dispersed and often only tackling a specific aspect of a policy domain, sometimes too generic. Following these standards is not usually mandatory.

Table 9 - Gap identification for N-T-1, Cope with the production of huge volumes of data

N-T-1	**Cope with the production of huge volumes of data**
Probably one of the biggest needs for administration is to keep up with the technical innovation. To cope with the production of huge volumes of data is a technical problem as well as a big challenge for the staff. On the one hand, there should be established technical infrastructure for new policies and the increasing number of data, on the other hand, the staff needs to be trained and able to manage data and produce "good" data. The interviewed division head in the policy domain "Youth and Welfare" on regional level stressed the importance of having enough staff that is able to handle data. To cope with the technical challenges, it is important that public administration is technical modernised and updated, which in turn requires financial investments. The automation of standardised processes could save a lot of time and resources (OECD 2013).	
Functionality N-T-1.F1	Technical infrastructure to support new policies and increasing amount of data.
Gap N-T-1.F1.G1	Some infrastructures have been deployed (like BDTI or MapR) and some specific tools are available as well, solving specific problems in the scope of public administrations. What it likes to be missing to further develop these capabilities in the public administrations is a tool set of infrastructures and tools and filling the gap between them and the real needs of public administrations, so these tools can be adapted to solve each specific problem.
Functionality N-T-1.F2	Staff training to be able manage and produce "good" data.
Gap N-T-1.F2.G2	Technical training is usually available for technical staff. It looks like public administrations management still has to be aware of the potential of using these infrastructures and tools to support policy making, while at the same time incorporating data scientists to PA staff.

Table 10 - Gap identification for N-T-3, Ensuring data security taking into account the protection of citizens' privacy

N-T-3	**Ensuring data security taking into account the protection of citizens' privacy**
Concerns about insufficient security and privacy are ubiquitous when it comes to the use of new technical possibilities - especially in public management (OECD 2017a). Besides the advantages and potentials, digitisation is associated with some technical and non-technical obstacles. Data protection and information security management can help to preserve trust in government[27]. Public administrations have to guarantee citizens´ informational self-determination, protect their sensitive personal data against unwarranted access and avoid unintended consequences (for example AI bias and identity theft). Additionally, it is necessary to ensure information security to managing sensitive information including people, processes and information systems. In this context an interviewee mentioned, that it is not even possible for them to send encrypted emails at present time.	
Functionality N-T-3.F1	Data protection, complying with GDPR standards for gathering, processing and storing personal data.
Gap N-T-3.F1.G1	How organizations obtain and use consent. Many individuals do not hold a digital footprint, so organizations need to be able to provide consent and consent management in hard copy as well as online.
Functionality N-T-3.F2	Information security management systems manage sensitive information so that it remains secure. It includes people, processes and IT systems by applying a risk management process.

[27] Website of the EU General Data Protection Regulation (GDPR), https://www.eugdpr.org/, retrieved February 7, 2018.

Gap N-T-3.F2.G2	To find a balance between the need for high security standards while ensuring enough openness to new innovation.
Gap N-T-3.F2.G3	The removal of explicit personal information from the citizens' data collected may not fully protect privacy, as combining multiple datasets may lead to the re-identification of individuals.

Table 11 - Gap identification for N-T-4, Establishment of a comprehensive technical infrastructure and IT architecture

N-T-4	**Establishment of a comprehensive technical infrastructure and IT architecture**
All interviewees stated that there is room for improvement in the technical infrastructure. The used technical infrastructure is partly outdated and does not meet current requirements, a fact that consequently increases administrative costs and leads to unnecessary bureaucracy. In addition, the lack of good infrastructure makes digitalisation difficult. In concrete terms, interface problems must be solved and harmonised. Concrete requirements that have been addressed in the various interviews are a comprehensive data infrastructure component, centralised records management and the ability to work mobile. This technical need is particularly related to the policy implementation and formulation but is also relevant in the other stages.	
Functionality N-T-4.F1	IT infrastructures are the backbone of a system of services that Public Administrations use and provide to citizens. They must be reliable, secure and economically sustainable.
Gap N-T-4.F1.G1	Decisions over the IT infrastructure are usually left to the initiative of each administration and undertaken without a shared vision, coordination or planning.
Gap N-T-4.F1.G2	IT Infrastructure managed with insufficient or fragmented resources in terms of budget but also skills, as acquisition of talent with specific capabilities to the different technical roles, is not fomented.
Gap N-T-4.F1.G3	Lack of technical capacity in policy processes and public administration own inability to understand technical data.

Table 12 - Gap identification for N-I-1, Link between impact, quality, performance measurements and financial information

N-I-1	**Link between impact, quality, performance measurements and financial information**
For making administrations not only more efficient but also more effective, activities and their costs should be closely linked to strategic outcomes and broader policy objectives. A monitoring with restricted focus on financial aspects in order to assess success of public services and political programmes is not enough. To reach a holistic view on success, it is more important to consider financial ratios interlinked with quality data, impact measurements and other performance indicators. For this reason, a strategic management system requires the integration of both financial and nonfinancial performance information. (Kaplan and Noerton 1992). This need was also validated by the focus group (policy domain "Social security"). As an example, one interviewee argued that there are missing linked information between the granted aid deliveries and the qualitative implementation by the institutions or care providers. Linking performance and outcome measurements to financial information provides information that is more relevant to decision makers (Pollanen 2016). The need is especially relevant in monitoring and evaluation, but also in the agenda setting and formulation stage. We were asking a PhD candidate with the thematic priority on digitalisation in the public sector in an interview, which experiences he gained during his work on use cases in the context of smart cities. He answered, that he was deeply shocked by the fact how low the information level of public administrations is, regarding their main tasks, services and societal outcomes. This leads to the point, that public administrations, before they want to link financial and nonfinancial information, quite	

generally first need to collect the relevant data (see also Need: Ensure availability of (real-time) information and knowledge), which can be integrated in a holistic interlinked monitoring system.	
Functionality N-I-1.F1	Strategic management system integrating both, financial and nonfinancial performance information. Financial ratios linked with quality data, impact measurement and other performance indicators.
Gap N-I-1.F1.G1	Although some specific challenges are presented from the public sector due to the variety of services, this is a usual procedure in the private sector. This requires the interlink of financial and non-financial indicators in a system to have an integral view of the performance and the financial effort required. It is likely to require a change in the mindset in the public administrations to set-up these procedures, as there are already tools and methods available to perform this.

Table 13 - Gap identification for N-I-3, Ensure availability of (real-time) information and knowledge

N-I-3	**Ensure availability of (real-time) information and knowledge**
Information is an asset that is constitutive to the effective and efficient supply of public services. To ensure that information meets the purposes for which it is intended, it must be accurate, accessible, valid, timely, complete and relevant (relevance especially means regional explicit information) (Hanger et al. 2013). In all the interviews that we conducted, it has become very clear and verified that information plays a very important role in policy making processes. According to the interviewed researcher in the field of administrative science, real-time data becomes relevant especially for the operative administration on the local level, for example, in the field of infrastructure. Information also plays an important role in economic policy. Up to now, current economic policy is based on very precise but outdated data. However, in such a dynamic environment, having up-to-date information is of great relevance. The interviewed division head in the policy domain “Youth and Welfare" emphasised that a good information situation, which means a certain amount of information in a good quality, is a precondition for further analyses and evaluations. In areas where there is already many data, initial success has been achieved. Nevertheless, there is still room for improvement here. However, it has been restricted that more than information is needed to positively change the policy process. Organisational conditions must be established to use this information adequately. For example, employees need to be able to understand and to use this information as well as to find creative solutions. This need seems to be closely connected with other needs, such as a comprehensive knowledge and information management, a deeper understanding IT potential and IT processes, and the establishment of a target-oriented personnel development.	
Functionality N-I-3.F1	Availability of accurate, accessible, valid, timely complete and relevant information.
Gap N-I-3.F1.G1	Regarding use cases, most of them goes form the most general to some specific areas, but not the most fundamental in terms of policy, like economics and taxes, social and welfare, that would have a real impact in the long term. Most of them just allow citizens to provide opinion or get information on secondary issues, or the information collected is just for accessory issues. They mostly provide the ability to manage limited impact areas. Most of the specific datasets and standards are generic catalogues, and the few specific are mostly about health-related issues, like food.
Functionality N-I-3.F2	Organisational conditions established to use the information adequately (employees need to be able to understand and use the information as well as to find creative solutions).
Gap N-I-3.F2.G2	There is a wide range of casuistic, but this mostly depends on the application and the type of data or tools. In general, there are a lot of experiences, and the main challenge here is to provide valuable information to the right decision-level inside public bodies, and easy to understand and meaningful information to citizens.

Table 14 - Gap identification for N-I-4, Comprehensive knowledge and information management

N-I-4	Comprehensive knowledge and information management
Knowledge management affects the organisation´s technical assets as well as the employees' willingness to share knowledge. Knowledge is an essential resource in public administrations and has to be stored in order to not get lost for the organisation. As a main reason for the loss of knowledge, participants of our focus group with a social political background named the retirement of employees. That is why it is important to build up a learning culture, to ensure and promote knowledge transfer within the organisation, as well as with relevant stakeholders (Hanger et al. 2013, OECD 2017b, Wige 2002).	
Functionality N-I-4.F1	Knowledge in the Public Sector should be collected, stored, shared and eventually destroyed.
Gap N-I-4.F1.G1	Rewards and learning & development processes in place in public administrations. The fear of not receiving recognition and accreditation from managers and colleagues can be the cause of retaining ownership. Besides, knowledge acquisition and high skilled and experienced staff is normally not a high priority.
Gap N-I-4.F1.G2	Sometimes IT infrastructure is old, so employees may lack the means and also the general skills of how to effectively share their knowledge.
Gap N-I-4.F1.G3	Lack of autonomy in the hierarchy and also lack of leadership and coordination, with a large number of professionals working in silos, which makes it difficult to share best practices.
Gap N-I-4.F1.G4	There is a lack of clear communication about the benefits and values of knowledge sharing.
Gap N-I-4.F1.G5	Some knowledge is hard to formalize as it is connected with the individual experience of a particular person.
Gap N-I-4.F1.G6	Insufficient capture, evaluation, feedback, communication and tolerance of past mistakes that would enhance information management in the public organizations.
Gap N-I-4.F1.G7	Loss of control over the location, distribution and use of knowledge due to the current facilities for generating, editing and storing documents. Public sector needs to break some mental schemes as they still think that power is in keeping it, rather than sharing it.

5 Research challenges on the use of big data for policy making

5.1 Research Clusters

We define six main research clusters related to the use of Big Data in policy making. Four of them are built on the Big Data cycle and value chain, while two are transversal at each phase of the cycle (Figure 12).

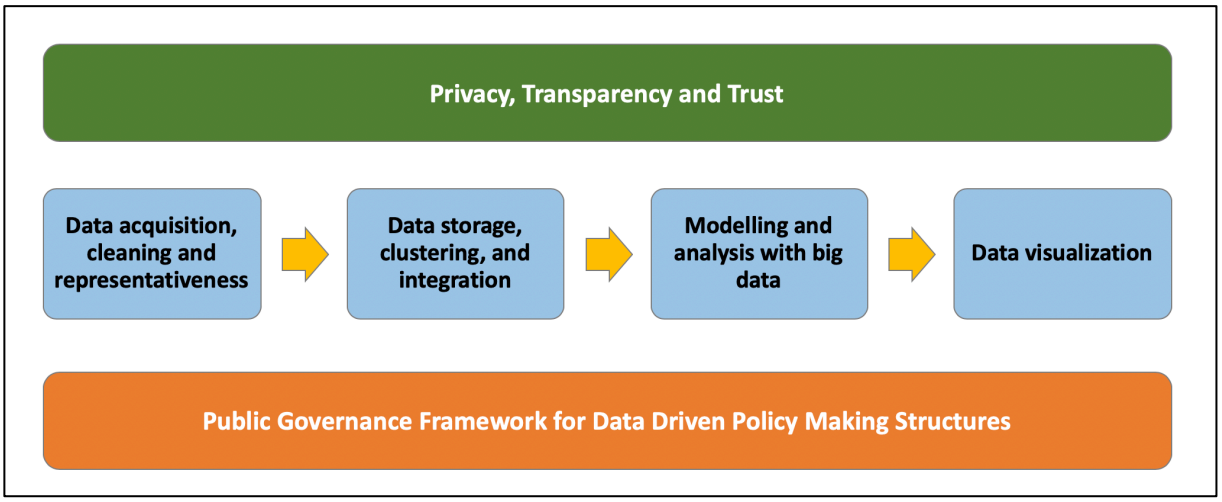

Figure 12 – Structure of the Research Clusters

Let us present know each cluster more thoroughly.

5.1.1 Cluster 1- Privacy, Transparency and Trust

This research cluster is transversal with respect to the others, and deals with core elements such as data ownership, security and privacy from one side, and transparency of the policy making on the other side. The overall aim is to increase trust on the government, especially on the public services, and a fair policy making activity and public service provisioning. A robust governance is crucial: even more than with traditional IT architectures, Big Data requires systems for determining and maintaining data ownership, data definitions, and data flows (inter a., Danaher et al. 2017). In fact, Big Data offers unprecedented opportunities to monitor processes that were previously invisible. In addition, the detail and volume of the data stored raises the stakes on issues such as data privacy and data sovereignty. Taking into account healthcare, developments such as crowdsourcing, participatory surveillance, and individuals pledging to become "data donors" and the "quantified self" movement (where citizens share data through mobile device-connected technologies), have great potential to contribute to our knowledge of disease, improving diagnostics, and delivery of healthcare and treatment (Kostkova et al. 2016). Therefore, there is the need to research the data regulation and standards for data generated by devices sensors or social media, identification frameworks to ensure ownership, privacy and security of personal data. In this regard a core objective should be to have a clear data privacy and security policies, and ensuring ownership of the data and regulations regarding the usage of the data generated by the devices or sensors. This is in order to avoid risks such as data usage for purposes other than providing the service,

inappropriate data storage and exposure of crucial personal data. The output of such research cluster includes a legal framework to ensure ownership, security and privacy of the data generated by the user while using the systems in the public administration. A second facet of this research cluster is transparency in the policy making process and availability of information and data from the public administration. Concerning the scrutiny of policy making creation, open data and public sector information allow a more generalized evaluation of the policies implemented and of their results. Moreover, publishing data leads to more transparency, new businesses, better evidence-based policy making and increased public sector efficiency only if the different actors in the chain have co-ownership of the data and be able to participate directly in its correction. In this sense free licensing and shared platform to publish and offer feedback/corrections directly to the data are crucial. Concerning the transparency in the policy making process, computer algorithms are widely employed throughout our economy and society to make decisions that have far-reaching impacts, including their applications for education, access to credit, healthcare, and employment. On the other side 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, unexpected behaviour of the system, and biased decision making (based on bias in the training data). Examples are given by AI techniques able to make predictions are based on huge data volumes (Centre for Public Impact, 2017). For instance, law enforcement agencies use AI technologies to predict areas where crimes are more likely to occur[28], or the use of algorithms for the automatic detection of fraudulent behaviour within government service provision (e.g. subsidies and social welfare). Other applications include prediction of criminal recidivism of the assessment of job applications, which have incurred in gender or racial discrimination. A final example is given by machine learning algorithms used for early detection of diseases, which can infringe the data protection rules on the use of non-anonymized medical records. Specifically, according to Mittelstadt et al. (2016), there are six main types of ethical concerns regarding algorithms. The first three are epistemic concerns: inconclusive evidence, inscrutable evidence and misguided evidence; then there are two normative concerns: unfair outcomes, transformative effects; and the last one, in traceability. Policymakers should therefore hold institutions using such analytics to the same standards as institutions where humans have traditionally made decisions and developers should plan and architect analytical systems to adhere to those standards when algorithms are used to make automated decisions or as input to decisions made by people. A crucial element, which is taking more and more importance in the last decade, is the practice of co-creating public services and public policies with citizens and companies, which would make public services more tailored to the needs of citizens and would open the black box of the inner working of public administration (inter al. Osborne et al. 2013). In the context of big data, co-creation activities take the form of citizen science-like activities such as data creation on the side of citizens, and in the co-creation of service in which disruptive technologies such as big data are adopted. An interesting research avenue that is gaining importance is the co-creation of the algorithms that are used in policy making, especially through serious games and simulations. Finally, openness and availability of government data for re-use provides the possibility to check and put under scrutiny the policy making activity (e.g. the UK-oriented initiative of My2050).

[28] For instance applications such as PredPol or CrimeScan used in various law enforcement agencies.

5.1.2 Cluster 2 - Public Governance Framework for Data Driven Policy Making Structures

The governance concept has been on the roll for the last couple of years. Universities have been founded, master study programmes have been established and it is one of the main topics in the horizon 2020 research programme of the European Commission. But what is the governance concept actually about? Plenty of different governance approaches and definitions can be found in the scientific community. Generally, the governance notion stands for shaping and designing areas of life in the way that rules are set and managed in order to guide policy-making and policy implementation. Core dimensions of governance are efficiency, transparency, participation and accountability (United Nations, 2007). Considering governance in the context of evidence-based policymaking, all dimensions are of utmost relevance. Corresponding to the definition of electronic governance, evidence-based and data informed policymaking in the information age applies technology in order to efficiently transform governments, their interactions with citizens and the relationship with citizens, businesses, other stakeholders, creating impact on the society (Estevez and Janowski, 2013). More concrete, digital technologies are applied for the processing of information and decision-making, the so called smart governance approach is applicable here (Pereira et al., 2018). Smart governance should support the policy makers in terms of time to devote to the policy making process; understanding the problems that need to be addressed, considering potential alternatives and the ability to identify the best solution (Bruce and Stiefel, 2012). In this frame, governance has to focus on how to leverage data for more efficient, rational, participative and transparent policy making. Although the governance discussion is not the newest one, it is a manifold challenge in the context of governmental and political responsibilities in the era of digital transformation.

5.1.3 Cluster 3 - Data Acquisition, Cleaning and Representativeness

Data to be used for policy making activity stem from a variety of sources: government administrative data, official statistics, user-generated web content (blogs, wikis, discussion forums, posts, chats, tweets, podcasting, pins, digital images, video, audio files, advertisements, etc.), search engine data, data gathered by connected people and devices (e.g. wearable technology, mobile devices, Internet of Things), tracking data (including GPS/geolocation data, traffic and other transport sensor data), and data sources collected through participation of citizens science activities. This leads to a huge amount of data that can be used and are of an increased size and resolution, span across time series, and that they are not, in most cases, collected by means of direct elicitation of people. While surveys, interviews, experiments, etc. require the active engagement of participants, most digital data are collected in the background. The advantage of this almost invisible footprint is a smaller likelihood of Hawthorne effect (inter al., Monahan and Fisher 2010;), in which individuals modify an aspect of their behaviour in response to their awareness of being observed or part of a study. There is another important consequence of the invisibility of digital data collection: digital data and their enhanced large version, big data, are well suited to capture behavioural information more than traditional social scientific instruments. However, concerning data quality, a common issue is balance between random and systematic errors. Random errors in measurements are caused by unknown and unpredictable changes in the measurement. In that regard, the unification of data so as to be editable and available for policy making is of extreme

importance: cancelling noise for instance is challenging. These changes may occur in the measuring instruments or in the environmental conditions. Normally random errors tend to be distributed according to a normal or Gaussian distribution. One consequence of this is that increasing the size of your data helps to reduce random errors. However, this is not the case of systematic errors, which are not random and therefore they affect measurements in one specific way. In this case, errors are from the way how data are created and therefore very large datasets might blind researchers to this kind of errors. Besides the potential presence of systematic errors, there two more methodological aspects of big data that require careful evaluation: the issue of representativeness and the construct validity problem (Veltri, 2019). Overall, for policy makers, the implications of these methodological considerations are that most big data will be a combination of existing and 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 of the results. Moreover, an important message is that although Big Data can greatly improve our understanding of socio-economic processes, they are not immune to error and biases(Bartlett, Lewis, Reyes-Galindo, & Stephens, 2018). It is impossible to screen out ambiguity and potential sources of systematic error. In other words, it is highly recommended that big data are not treated as a 'magic bullet' that can provide answers to all social and economic problems. Therefore, the appropriateness of any Big Data source for decision-making should be made clear to users. For this reason, 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. The ideal would be a cleansing mechanism for reducing the inaccuracy of the data to the smallest extent, though, especially in case this can be predicted beforehand.

5.1.4 Cluster 4 - Data Storage, Clustering, and Integration

This research cluster deals with information extraction from unstructured, multimodal data, heterogeneous, complex, or dynamic data. Heterogeneity and incomplete data must be structured prior to the analysis in a homogeneous way, as most computer systems work better if multiple items are stored in an identical size and structure. But an efficient representation, access and analysis of semi-structured data is necessary because as a less structured design is more useful for certain analysis and purposes. Specifically, the large majority of big data, from the most common such as social media and search engines data to transactions at self-check out in hotels or supermarkets, are generated for different and specific purposes. They are not the design of a researcher that elicits their collection with in mind already an idea of a theoretical framework of reference and of an analytical strategy. Specifically regarding data from social media, they can be really challenging to clean and demand a lot of effort. What is more, the data elicited from social media could be biased. Big data, by contrast, just are a large universe of such correlations—very often they are not carefully designed. Twitter and big national surveys have been both uses to analyse public opinion but their data are different and so it is different what they can reveal about public opinion. Sentiment analysis on Twitter data, the emotional valence of tweets computed by text mining, is now a popular way of tracking public opinion mood and not well suited for surveys. From this point of view, the debate about big data enthusiasts and sceptics should be formulated differently. There are research questions and issues for which big data are interesting and other for which 'traditional' social scientific methods are still more reliable and useful.

Therefore, one of the first characteristics of big data, highly relevant for the social scientist is their 'organic' nature in contrast with 'designed' (for social research data). Currently data are becoming a

cheap commodity around, simply because the society has created systems that automatically track transactions of all sorts. For example, Internet search engines build data sets with every entry, Twitter generates tweet data continuously, traffic cameras digitally count cars, scanners record purchases, Internet sites capture and store mouse clicks. Collectively, human society is assembling data on massive amounts of its behaviours. If we think of these processes as an ecosystem, it is self-measuring in increasingly broad scope. Indeed, we might label these data as 'organic', a now-natural feature of this ecosystem. Therefore, big data are considered 'organic', they are created by different actors in the context of producing or delivering goods or services and not for research. In this respect, common to big data is the idea of the repurposing of data. Data that were collected for other initial aims are repurposed for new specific research goals set by the secondary analyst. The difference is that for big data, especially those collected by private companies, the lack of transparency about how data are collected or coded is a problem that has to be faced.

Repurposing of data requires a good understanding of the context in which the data repurposed were generated in the first place. In other words, these are not 'natural', they are the outcome of designers and socio-economic processes, therefore created with some goals and trade-offs. It is about finding a balance between identifying the weaknesses of the repurposed data and at the same time finding their strengths. In synthesis, the combination and meaning extraction of big data stemming from different data sources to be repurposed for another goal 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. Further to the identification of patterns, trends and relevant observables, and extraction of relevant information and feature extraction from heterogeneous databases, there is the need to ensure interoperability and exchange of data and information from different databases within the public administration. Finally, and very importantly, a pre-requisite of clustering and integration is the presence of tools and methodologies to successfully store and process big data.

5.1.5 Cluster 5 - Modelling and Analysis with Big Data

The intrinsic complexity of the emerging challenges human beings collectively face requires a deep comprehension of the underlying phenomena in order to plan effective strategies and sustainable solutions: from the planning of urban infrastructures to containment strategies for pandemics, from the impact of political campaigns to measures against information pollution and misinformation. In this regard, a main challenge in the use of big data for applications related to policy making is copying with unanticipated knowledge. One of the key problems when forecasting is represented by a lack of knowledge about what could be, i.e., about that peculiar space where lie everything that is not yet actual, still possible, the so-called space of the possible. In this framework, a beautiful notion is that of the "adjacent possible". Originally introduced in the framework of biology, the adjacent possible metaphor already expanded its scope to include all those things (ideas, linguistic structures, concepts, molecules, genomes, technological artefacts, etc.) that are one step away from what actually exists, and hence can arise from incremental modifications and recombination of existing material. The strange and beautiful truth about the adjacent possible is that its boundaries grow as one explores them. Unfortunately, we are very bad at grasping this space. There is a good reason why we are very bad at conceiving the way in which we explore this space. We are trying to conceive the occurrence of something new, something that never occurred before. The term "Unanticipated Knowledge" refers precisely to the observation of events whose existence cannot even been foreseen. One typical solution is looking at the future with

the eyes of the past. This means looking at the time series of past events, hoping that this is enough to predict the future. We know this is not working. This was the first attempt, for instance, for weather forecast. And it failed, because of the great complexity of the underlying phenomenon. We now know that predictions have to be based on modelling, which means constructing a model of the phenomenon, possibly driven by relevant sets of data, and simulating it, projecting the system into the future. The availability of huge amounts of data could certainly help in this direction, though it does not represent per se a general solution. The point is that data (also big data) tell us something about the past and the knowledge of the past is not always helpful in designing the future. Looking at the future with the eyes of the past could be misleading also for machines. Despite the recent dramatic boost of inference methods, they still crucially rely on the exploitation of prior knowledge and the problem of how those systems could handle unanticipated knowledge remains a great challenge. In addition, also with the present available architectures (feed-forward and recurrent networks, topological maps, etc.) it is difficult to go much further than a black-box approach and the understanding of the extraordinary effectiveness of these tools is far from being elucidated. Given the above-mentioned context it is important to make steps towards a deeper insight about the emergence of the new and its regularities. This implies conceiving better modelling schemes, possibly data-driven, to better grasp the complexity of the challenges in front of us, and aiming at gathering better data more than big data, and wisely blending modelling schemes. But we should also go one step further in developing tools allowing policy makers to have meaningful representations of the present situations along with accurate simulation engines to generate and evaluate future scenarios. 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. In short scenario simulators and decision support tools. In this framework it is highly important to launch new research directions aimed at developing 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. Taking into account the development of new models, there are basically two main approaches (Kim et al. 2017): data modelling and simulation modelling. Data modelling is a method in which a model represents correlation relationships between one set of data and the other set of data. On the other hand, simulation modelling is a more classical, but more powerful, method in which a model represents causal relationships between a set of controlled inputs and corresponding outputs. Clearly data modelling suffers some limitations, such as the inability to predict under changed conditions, as well as the inability to cope with unexpected events. On the other hand, the simulation model has the following property: if knowledge about the system can be obtained, it should be applied to the prediction. In addition, the simulation model requires idealistic assumptions and constraints about the system, while the data model does not.

5.1.6 Cluster 6 - Data Visualization

Implementing effective data visualization solutions for Big Data has to take into account, apart the volume of the data, other intrinsic constraints generated by the typical characteristics of Big Data: real-time changes, extreme variety of the sources, and different levels of data structuring. Specifically, 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 of the results of big data

modelling and analysis. Specifically, the explosion in computing techniques led to the generation of a tremendous amount of data which are stored in the cloud and processed in the IT infrastructures all over the world.[29] In managing this huge amount of data, when it comes to human-computer interaction there is a need to distil the most important information to be presented it in a humanly understandable and comprehensive way. Here it comes visualisation, which is a way to interpret and translate data from computer understandable formats to human ones by employing graphical models, charts, graphs and other images that are conventional for humans. From one hand we can define visualization as any technique for creating insight, preferably by allowing users to interact and alter with the visualization to iteratively solve questions and form new questions based on previous findings. On the other hand, visualization can be defined as a set of techniques for communicating knowledge that can be supported by data. In contrast with visualization traditionally seen as the output of the analytical process, visual analytics considers visualization as a dynamic tool that aims at integrating the outstanding capabilities of humans in terms of visual information exploration and the enormous processing power of computers to form a powerful knowledge discovery environment. In this view visual analytics is useful for tackling the increasing amount of data available, and for using in the best way the information contained in the data itself. Moreover, visual analytics aims at present the data in way suitable for informing the policy making process. More in particular the interdisciplinary field of visual analytics aims at combining human perception and computing power in order to solve the information overload problem. Visualisation and visual analytics should be considered in strict integration with other research areas, such as modelling and simulation, social network analysis, participatory sensing, open linked data, visual computing. With regard to the governance and policy making context, some visualization tools can be applicable to a wide array of issues and situation (education, environment, public health, urban growth, national defence, etc.). In the public context, visual analytics of public data is an exploding field, with particular relation to the open data movement, in order to monitor policy context and evaluate government policies. Today's governments face the challenge of understanding an increasingly complex and interdependent world, and the fast pace of change and increased instability in all the areas of regulation requires rapid decision making able to draw on the wider amount of available evidence in real-time. How can visualization and visual analytics help? First, generate high involvement of citizens in policy-making. One of the main applications of visualization is in making sense of large datasets and identifying key variables and causal relationships in a non-technical way. Similarly, it enables non-technical users to make sense of data and interact with them (Vornhagen et al. 2018). Further, good visualization is also important in "selling" the data-driven policy making approach. Policy makers need to be convinced that data-driven policy making is sound, and that its conclusions can be effectively communicated to other stakeholders of the policy process. External stakeholders also need to be convinced to trust, or at least, consider data-driven policy-making. There should be a clear and explicit distinction of the audiences for the policy visualisations: e.g. experts, decision makers, the general public. Experts are analyzing data, are very familiar with the problem domain and will generate draft policies or conclusions leading to policies Decision makers may not be technical users, and may not have the time to delve deep into a problem. They will listen to experts and must be able to understand the issues, make informed decisions and explain why. The public needs to understand the basics of the issue and the resulting policy in a clear manner. A second element is that visualization help to understand the impact of policies: visualization is instrumental in making evaluation of policy impact more effective. Finally, it helps to identify problems at an early stage, detect the "unknown unknown" and

[29] For an overview of Big Data sources, please refer to https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP7_Big_data_sources_overview1

anticipate crisis: visual analytics are largely used in the business intelligence community because they help exploiting the human capacity to detect unexpected patterns and connections between data. Thereby they help early detection of potential threats at an early stage. Considering specifically Big Data visualization, it has to be taken into account, apart from the sheer volume of the data, other intrinsic constraints generated by the typical characteristics of Big Data are changes in real-time, extreme variety of the sources, and different levels of data structuring. In this respect, it is better to use several visualization techniques simultaneously to better illustrate relationships among a large amount of data. Finally, data visualization can play a specific role in several phases of the Big Data Life Cycle: in the pre-processing, staging, handling phase; in exploratory data analysis, and in presentation of analytical results.

There are three main visualization instruments for Big Data:

- Infographic and information design: the art and science of preparing and presenting the information so that they can be used by humans in an efficient and effective;
- Visual analytics: graphic techniques to analyze and make sense of the data;
- Dashboards: graphic techniques to measure and monitor relevant data of an organization, in order to achieve their fixed objectives.

Similarly, the visual analytics techniques adopted for Big Data are:

- Visual Analytics techniques used to extract meaningful patterns, outliers, clusters and gaps;
- Interactive visualization used to discover the most interesting relationships among data, investigate what-if scenarios, verify the presence of biases; simulate the impact of changes;
- Dissemination tools, used to enlighten the sense of data and tell stories about them.

5.2 From Research Gap to Research Clusters

Clearly there is a conceptual correspondence between the research gaps, needs and the research clusters. In this respect, stemming from the gaps identified in section 4, we developed a series of research needs mapped against the research clusters in Table 15.

Table 15 – Mapping of research needs and clusters of research challenges

Research Need	Research Clusters					
	C1	C2	C3	C4	C5	C6
N-S-1: development of new evaluation frameworks and tools for the assessment of the impact of policies. Such evaluation frameworks should build on a set of evaluation criteria and indicators adapted to the specific domains	X	X				X
N-I-1: development of new procedures and tools for the establishment of a management system integrating both, financial and nonfinancial performance information linked with quality data, impact measurement and other performance indicators	X	X				X
N-S-2: development of new tools, methodologies and regulatory frameworks to boost participation of citizens in policies making by mean of crowdsourcing and co-creation of policies. In this regard, a way needs to be found to integrate impact assessment and sentiment analysis tools and techniques to gauge citizen opinion expressed via social media channels.	X	X	X			

N-S-4: development of new regulations, tools and technical frameworks that ensure absence of bias and transparency in the policy making process and cybersecurity of IT systems in the public administration	X					
N-S-9, N-O-7, N-I-3: development and deployment of frameworks and tools that allow the secure sharing of information and data within the public administration, as well as the interoperability of systems and databases. These frameworks include the standardization of organizational processes. In this regard, achieving near-complete interoperability across public systems and databases, together with the streamlining of organisational processes, are important prerequisites of technology acceptance.		X		X		
N-T-1: development of specific interoperable cloud infrastructures and (re-usable and integrating) models for the management and analysis of huge volumes of data		X		X	X	
N-T-3: development of new regulations, tools and technical frameworks that ensure absence respect of citizens' privacy and data ownership/security, especially in case the personal information need to be migrated across public administration agencies	X					
N-T-4: development and establishment of a unique reliable, secure and economically sustainable technical and IT infrastructure which would work as a backbone for all the public services developed and implemented in the public sector	X			X	X	
N-I-4: development of information management systems and procedures for the collection, storing, sharing, standardization and classification for information pertaining to the public sector	X	X		X	X	

5.3 Research Challenges

This final step deals with the presentation of the research challenges per each cluster. In the final version of the roadmap we will include a more extended presentation of the research challenges, as well as in particular the short and long term timeline for research. A schematic representation of the research clusters and related research challenges is provided in Table 16.

Table 16 – Research clusters and related research challenges

Research Cluster	Research Challenges
C1- Privacy, Transparency and Trust	RC 1.1 - Big Data nudging
	RC 1.2 - Algorithmic bias and transparency
	RC 1.3 - Open Government Datasets
	RC 1.4 – Manipulation of statements and misinformation
C2 - Public Governance Framework for Data	RC 2.1 - Forming of societal and political will
	RC 2.2 - Stakeholder/Data-producer-oriented Governance approaches

Driven Policy Making Structures	RC 2.3 - Governance administrative levels and jurisdictional silos
	RC 2.4 - Education and personnel development in data sciences
C3 - Data acquisition, cleaning and representativeness	RC 3.1 – Real time big data collection and production
	RC 3.2 - Quality assessment, data cleaning and formatting
	RC 3.3 - Representativeness of data collected
C4 - Data storage, clustering and integration	RC 4.1 - Big Data storage and processing
	RC 4.2 - Identification of patterns, trends and relevant observables
	RC 4.3 - Extraction of relevant information and feature extraction
C5 - Modelling and analysis with big data	RC 5.1 – Identification, acceptance and validation of suitable modelling schemes inferred from existing data
	RC 5.2 - Collaborative model simulations and scenarios generation
	RC 5.3 - Integration and re-use of modelling schemes
C6 - Data visualization	RC 6.1 – Automated visualization of dynamic data in real time
	RC 6.2 - Interactive data visualization

5.3.1 Research Challenges on Privacy, Transparency and Trust

Research Challenge 1.1 - Big Data nudging

Description. Nudging has long been recognized as a powerful tool to achieve policy goals by inducing changes in citizens behaviour, while at the same time presenting risks in terms of respect of individual freedom. Nudging can help governments, for instance, reducing carbon emissions by changing how citizens commute, using data from public and private sources. But it is not clear to what extent can government use these methods without infringing citizens' freedom of choice. And it is possible to imagine a wide array of malevolent applications by governments with a more pliable definition of human rights. The recent case of Cambridge Analytica acts as a powerful reminder of the threats deriving from the combination of big data with behavioural science. These benefits and the risks are multiplied by the combination of nudging with big data analytics, becoming a mode of design-based regulation based on algorithmic decision-guidance techniques. When nudging can exploit thousands of data points on any individual, based on data held by governments but also from private sources, the effectiveness of such measures – for good and for bad – are exponentially higher. Unlike the static nudges, Big Data analytic nudges (also called hypernudging) are extremely powerful due to their continuously updated, dynamic and pervasive nature, working through algorithmic analysis of data streams from multiple sources offering predictive insights concerning habits, preferences and interests of targeted individuals. In this respect, as pointed out by Yeung (2016), by "highlighting correlations between data items that would not otherwise be observable, these techniques are being used to shape the informational choice context in which individual decision-making occurs, with the aim of channeling attention and decision-making in directions preferred by the 'choice architect". In this respect, these techniques constitute a 'soft' form of design-based control, and it remains unchartered territory the definition of the scope, limitations and safeguards – both technological and not – to ensure the simultaneous achievement of fundamental policy goals with respect of basic human rights.

Relevance and applications in policy making. Behavioural change is today a fundamental policy tools across all policy priorities. The great challenges of our time, from climate change to increased inequality to healthy living can only be addressed by the concerted effort of all stakeholders. But in the present context of declining trust in public institutions and recent awareness of the risk of big data for individual freedoms, any intervention towards greater usage of personal data should be treated with enormous care and appropriate safeguards should be developed. Notwithstanding the big role of the GDPR, the trust factor is not understood well so far. While there are a number of studies on trust and there exist several trust models explaining trust relations and enabling empirical research on the level of trust, these researches are not yet including the study of trust in big data applications and the impact this may have on human behaviour. In this regard, there is the need to assess power and legitimacy of hypernudging to feed real-time policy modelling to inform changes in institutional settings and governance mechanisms, to understand how address key societal challenges exploiting the potential of digital technologies and its impact on institutions and individual and collective behaviours, as well as to anticipate emerging risks and new threats deriving from digital transformation and changes in governance and society.

Technologies, tools and methodologies. This research challenge stems from the combination of machine learning algorithms and behavioural science. Machine learning algorithms can be modelled to find patterns in very large datasets. These algorithms consolidate information and adapt to become increasingly sophisticated and accurate, allowing them to learn automatically without being explicitly programmed. At the same time, potential safeguards deal with transparency tools to ensure adequate consent by the citizens to be involved in such initiatives, as well as algorithm evaluation mechanisms for potential downside.

Research Challenge 1.2 - Algorithmic bias and transparency

Definition. Many decisions, are today automated and performed by algorithms. Predictive algorithms have been used since 20 years in public services, whether for predicting risks of hospital admissions or recidivism in criminal justice. Newer ones could predict exam results or job outcomes or help regulators predict patterns of infraction. It's useful to be able to make violence risk assessments when a call comes into the police, or to make risk assessments of buildings. Health is already being transformed by much better detection of illness, for example, in blood or eye tests. Algorithms are designed by humans, and increasingly learn by observing human behaviour through data, therefore they tend to adopt the biases of their developers and of society as a whole. As such, algorithmic decision making can reinforce the prejudice and the bias of the data it is fed with, ultimately compromising the basic human rights such as fair process. Bias is typically not written in the code, but developed through machine learning based on data. For this reason, it is particularly difficult to detect bias, and can be done only through ex-post auditing and simulation rather than ex-ante analysis of the code. There is a need for common practice and tools to controlling data quality, bias and transparency in algorithms. Furthermore, as required by GDPR, there is a need for ways to explain machine decisions in human format. Furthermore, the risk of manipulation of data should be considered as well, which may lead to ethical misconduct. In this regard, Zarsky (2016) provides a taxonomy of objections to algorithmic decision-making.

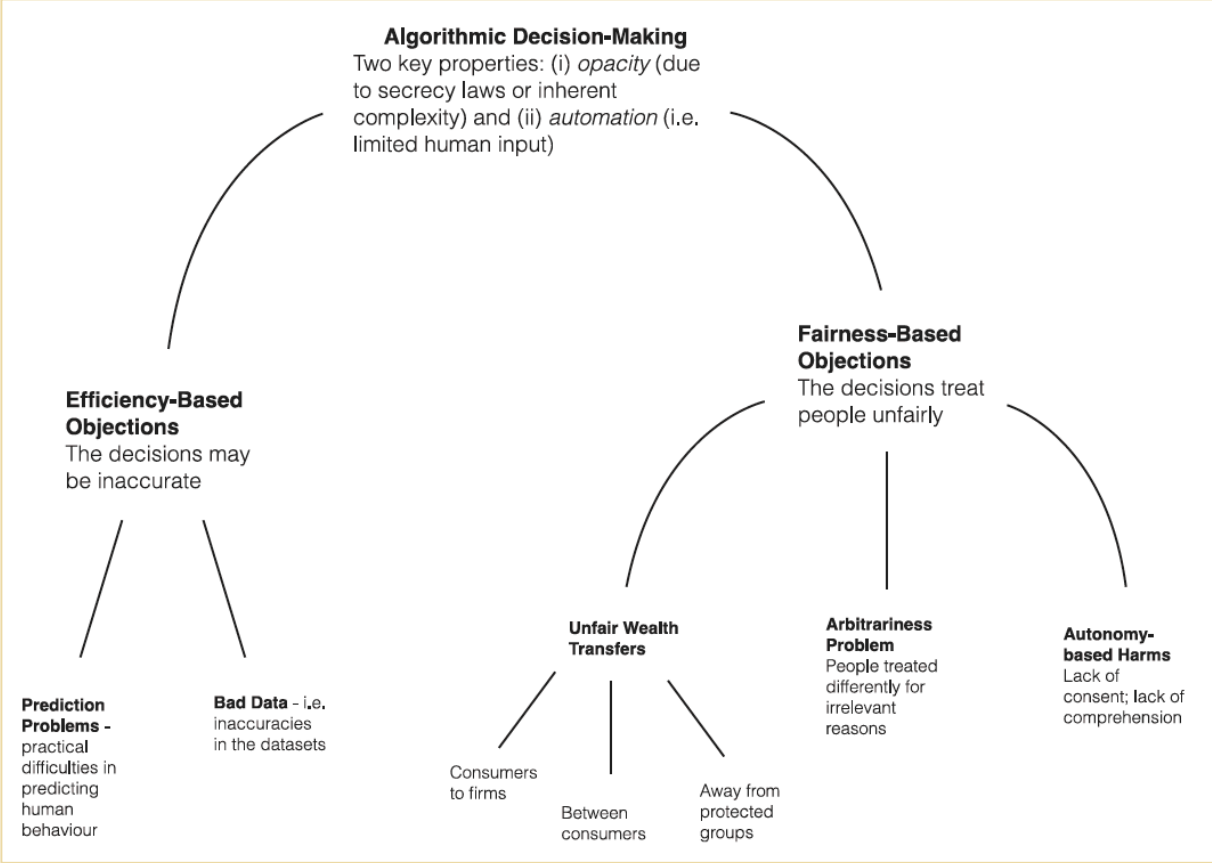

Figure 13 - Taxonomy of objections to algorithmic decision-making

Relevance and applications in policy making. Algorithms are increasingly used to take policy decisions that are potentially life changing, and therefore they must be transparent and accountable. GDPR sets out the clear framework for consent and transparency. Transparency is required for both data and algorithm, but as bias is difficult to detect in the algorithm itself and ultimately it is only through assessment of real-life cases that discrimination is detectable.

An interesting depiction of applications is provided by Engin and Treleaven (2019) in Figure 14.

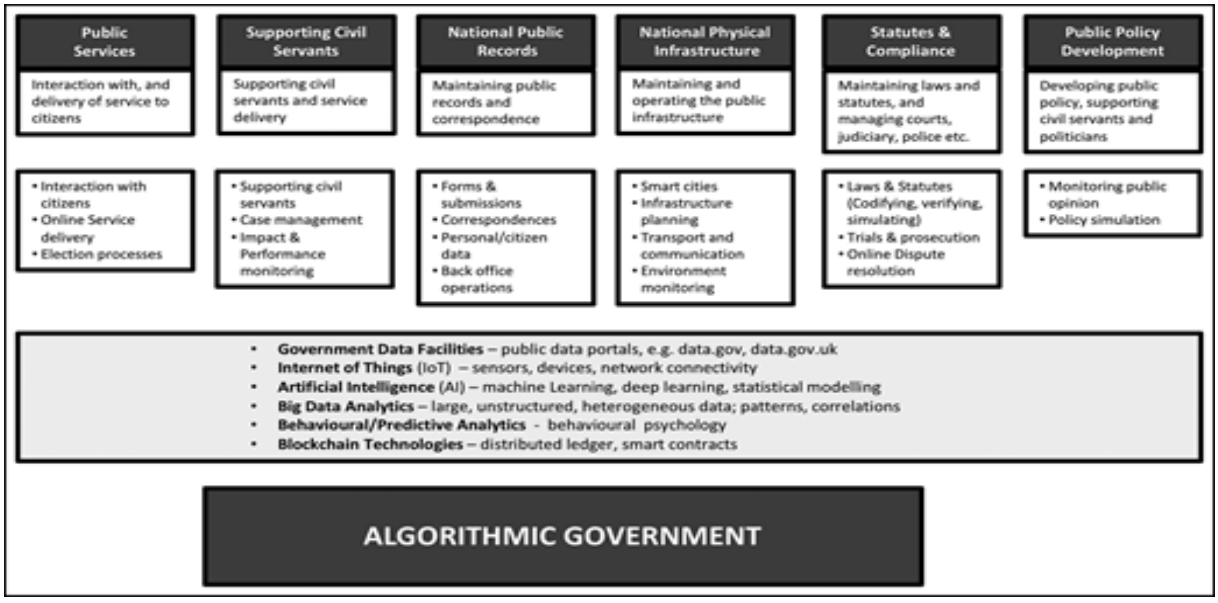

Figure 14 – Automation of Government Services (Source: Engin and Treleaven 2019)

Technologies, tools and methodologies. The main relevant methodologies are algorithm co-creation, regulatory technologies, auditability of algorithms, online experiments, data management processing algorithms and data quality governance approaches. Regarding governance, the ACM U.S. Public Policy Council (USACM) released a statement and a list of seven principles aimed at addressing potential harmful bias of algorithmic solutions: [30]

1. Awareness: Owners, designers, builders, users, and other stakeholders of analytic systems should be aware of the possible biases involved in their design, implementation, and use and the potential harm that biases can cause to individuals and society;
2. Access and redress: Regulators should encourage the adoption of mechanisms that enable questioning and redress for individuals and groups that are adversely affected by algorithmically informed decisions;
3. Accountability: Institutions should be held responsible for decisions made by the algorithms that they use, even if it is not feasible to explain in detail how the algorithms produce their results;
4. Explanation: Systems and institutions that use algorithmic decision-making are encouraged to produce explanations regarding both the procedures followed by the algorithm and the specific decisions that are made. This is particularly important in public policy contexts;
5. Data Provenance: A description of the way in which the training data was collected should be maintained by the builders of the algorithms, accompanied by an exploration of the potential biases induced by the human or algorithmic data-gathering process. Public scrutiny of the data provides maximum opportunity for corrections;
6. Auditability: Models, algorithms, data, and decisions should be recorded so that they can be audited in cases where harm is suspected;
7. Validation and Testing: Institutions should use rigorous methods to validate their models and document those methods and results. In particular, they should routinely

[30] For more information please refer to https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

perform tests to assess and determine whether the model generates discriminatory harm. Institutions are encouraged to make the results of such tests public.

Further, Geoff Mulgan from NESTA has developed a set of guidelines according to which governments can better keep up with fast-changing industries.[31] Similarly, Eddie Copeland from NESTA has developed a "Code of Standards for Public Sector Algorithmic Decision Making."[32] A very interesting initiative is also the one carried out by the Cyprus Center for Algorithmic Transparency, which is a new research center hosted at the Open University of Cyprus, with the mission to raise awareness about algorithmic biases, particularly in information access systems, and to develop interventions and tools to promote algorithmic transparency.

Research perspectives for Research Challenges 1.1 and 1.2. In this case, the research perspectives are common to the two research challenges.

A first research strand concerns the ethical implication and transparency of algorithms. For instance, Martin (2018) identifies a responsibility of the developers for their algorithms later in use, what those firms are responsible for, and the normative grounding for that responsibility. In his framework, algorithms are value-laden as they create moral consequences, reinforce or undercut ethical principles, and enable or diminish stakeholder rights and dignity. Further, according to him algorithms are an important actor in ethical decisions and influence the delegation of roles and responsibilities within these decisions. In this respect, his conclusion is that if an algorithm is designed to preclude individuals from taking responsibility within a decision, then the designer of the algorithm should be held accountable for the ethical implications of the algorithm in use. More in depth, Mittelstadt et al. (2016) provide a map of the ethics of algorithms, depicting a prescriptive framework of types of issues arising from algorithms (Figure 15).

[31] For more information please refer to https://www.nesta.org.uk/blog/anticipatory-regulation-10-ways-governments-can-better-keep-up-with-fast-changing-industries/

[32] For more information please refer to https://www.nesta.org.uk/blog/10-principles-for-public-sector-use-of-algorithmic-decision-making/

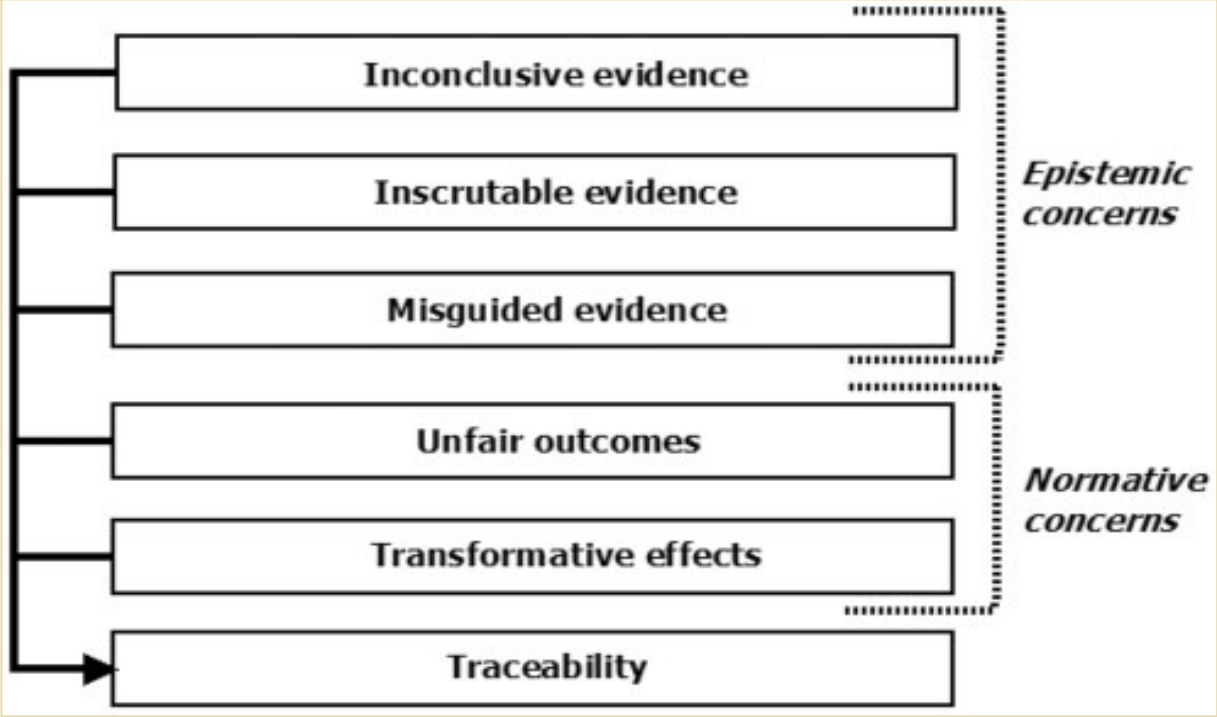

Figure 15 - Six types of ethical concerns raised by algorithms (Source: Mittelstadt et al. (2016)

Mittelstadt et al. (2016) identify also other research avenues, such as regarding how privacy operates at group level, absent of identifiability (e.g. Mittelstadt and Floridi, 2016; Taylor et al., 2017), mechanisms for enforcing privacy in data analytics (Agrawal and Srikant, 2000; Fule and Roddick, 2004), discrimination detection in data mining (e.g. Barocas, 2014; Calders and Verwer, 2010; Hajian et al., 2012), capacity of algorithms to disadvantage users in ways exceed the legal definitions of discrimination (Sandvig et al., 2014; Tufekci, 2015). Further research is required concerning shared responsibility across a network of human and algorithmic actors simultaneously (Simon, 2015), as well as de-responsibilisation of human actors (Davis et al., 2013; Zarsky, 2016). Further research is also required concerning malfunctioning (Floridi et al., 2014, Burrell 2016) or harmful actions and feedback loops (Orseau and Armstrong, 2016). A final domain concerns the operationalization of transparency, for instance requirements for algorithms to be explainable or interpretable (Tutt, 2016), algorithmic auditing carried out by external regulators (Pasquale, 2015; Tutt, 2016; Zarsky, 2016), data processors (Zarsky, 2016), or empirical researchers (Kitchin, 2016; Neyland, 2016), using reporting mechanisms designed into the algorithm itself (Vellido et al., 2012), or ex post audit studies (Adler et al., 2016; Diakopoulos, 2015; Kitchin, 2016; Romei and Ruggieri, 2014; Sandvig et al., 2014). Further research is also need in designing low impact auditing mechanisms for algorithms (Sandvig et al., 2014) based on transparency and interpretability of machine learning (e.g. Kim et al., 2015; Lou et al., 2013).
Interesting research is also carried out in the realm of algorithmic governance, where for instance Danaher et al. (2017) provide a taxonomy of the main research themes concerning legitimate and effective algorithmic governance (see Figure 16).

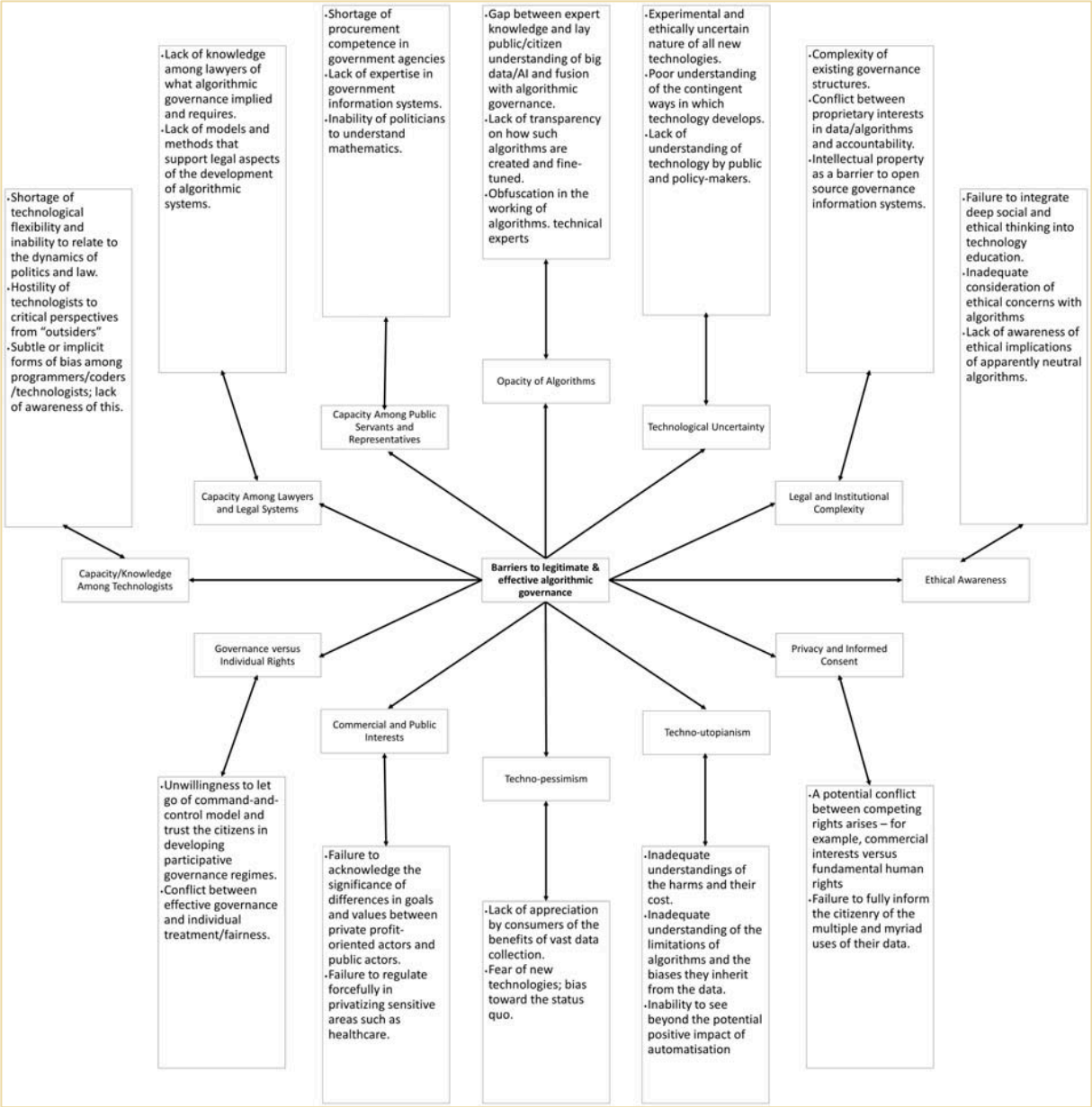

Figure 16 - Key research themes in response to barriers to legitimate and effective algorithmic governance (Source: Danaher et al. 2017)

Another strand of research concerns the development of impact assessment frameworks for algorithms. As an example, Engin and Koshiyama (2019) based on five main building blocks:

- Principles and Values;
- Policy & Law – Code of Conduct;
- Abstraction – Maths and Tech Formulation;
- Implementation and Technology;
- Compliance and Regulation.

Further, they develop an Artificial Intelligence impact assessment canvas (see Figure 17).

Decisions How these predictions will affect the end-user? What are the implications of these decisions to the end-user when the system's predictions are wrong?	**AI Assessment Task** Type of assessment, Group variables, Output to explain, Worst-case analysis.	**Safety propositions** What are we trying to do for the end-user(s) of the system? What objectives are we serving? How our decisions will affect the end-user(s) prospects?.	**Data Sources** Which raw data sources are we using? Are we infringing individual/company rights of privacy and agency in using these data sources? (internal and external)	**Collecting Data** Are the subject aware of the data collection? What are the access/control of the subjective over its data? The data being collected are actually being used during the modelling process, or are we just storing it?
Making Assessments When do we make on new predictions/decisions? How long do we have to wait to see the implications of the system's decisions?	**Offline assessment** Methods and metrics to evaluate the system before deployment, such as on Fairness, Explainability and Robustness.		**Features** Which demographics are we including/transforming? Are we paying attention specifically to vulnerable ones (e.g., children, minorities, disabled persons, elderly persons, or immigrants)? Are we applying transformations to some features that render them difficult/unable to be explained later on?	**Building Models** What are the limitations of the model being created/updated (Robustness, Explainability and Fairness)? What are the measures put in place to mitigate or eliminate these limitations are being addressed?
	Live Assessment and Monitoring Methods and metrics to assess the system after deployment, such as to quantify its reliability, fairness and interpretability during its decision-making.			

Figure 17 – Artificial intelligence impact assessment canvas (Source: Engin and Koshiyama 2019)

Similarly, Tal et al. 2019 provide a framework for the detection and reducing biases in algorithmic systems (see Figure 18).

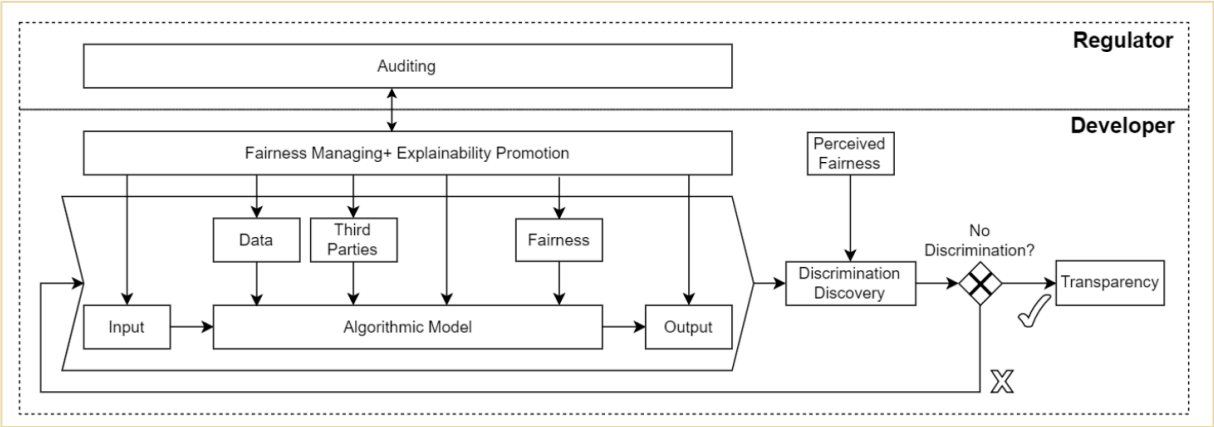

Figure 18 – Framework prototype (Source: Tal et al. 2019)

Another interesting strand is the one of privacy enhancing technologies: for instance McCarthy and Fourniol (2019) investigate the potential of such technologies in enabling governments to unlock the value of data, as well as the contingent and in principle limitations on the role of such technologies in ensuring well-governed use of data. Specific technologies investigated include Homomorphic encryption schemes, trusted execution environments, secure multiparty computation, differential privacy, and personal data stores. A final research strand that is being developed is the one of co-creation

of algorithms with citizens and business, through tools and methodologies such as online platforms and serious games.

Research Challenge 1.3 - Open Government Data

Definition. Open Data are defined as data which is accessible with minimal or no cost, without limitations as to user identity or intent. Therefore, this means that data should be available online in a digital, machine readable format. Specifically, the notion of Open Government Data concerns all the information that governmental bodies produce, collect or pay for. This could include geographical data, statistics, meteorological data, data from publicly funded research projects, traffic and health data. In this respect the definition of Open Public Data is applicable when that data can be readily and easily consulted and re-used by anyone with access to a computer. In the European Commission's view 'readily accessible' means much more than the mere absence of a restriction of access to the public. Data openness has resulted in some applications in the commercial field, but by far the most relevant applications are created in the context of government data repositories. With regard to linked data in particular, most research is being undertaken in other application domains such as medicine. Government starts to play a leading role towards a web of data. However, current research in the field of open and linked data for government is limited. This is all the more true if we take into account Big Data alimented by automatically collected databases. Further, two issues requiring greater understanding present themselves: first, is the nature of data as a dual purpose commodity with both economic and social value. Second, is the nature of the incentives available to encourage data providers to share their data for public benefit (Virkar et al. 2019). An important aspect is also the risk of personal data included in open government data or personal data being retrieved from the combination of open data sets. The Open Government data principles were defined in December 2007, during an Open Government Working Group Meeting held in Sebastopol (United States), which gathered 30 open government advocates:[33]

1. Complete: All public data are made available. Public data are data that is not subject to valid privacy, security or privilege limitations;
2. Primary: Data are as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms;
3. Timely: Data are made available as quickly as necessary to preserve the value of the data.
4. Accessible: Data are available to the widest range of users for the widest range of purposes.
5. Machine processable: Data are reasonably structured to allow automated processing.
6. Non-discriminatory: Data are available to anyone, with no requirement of registration.
7. Non-proprietary: Data are available in a format over which no entity has exclusive control.
8. License-free: Data are not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.

Finally, Ubaldi (2013) provides a set of conditions for availability and re-usability of open government data:

- Availability and accessibility:
    - Data are easily accessible, e.g. it is available in disaggregated forms and in electronic format, and the right to access data in electronic format is recognized;
    - Data are available in a convenient and modifiable form;

[33] https://public.resource.org/8_principles.html

    - Data are easy discoverable and findable;
- Re-use and distribution
    - Data are in machine-readable format;
    - Data are released in open formats (specifications have been made public and there is no need of having a specific software to use the information) which are machine-readable;
    - Data are available through bulk downloads thus enabling access not just to one or two pieces of government data, but to full datasets;
    - Data are linked and released in a timely fashion;
    - Users have the right to re-use data without discrimination.

Relevance and applications in policy making. Clearly opening government data can help in displaying the full economic and social impact of information, and create services based on all the information available (e.g. Virkar et al. 2018). Other core elements in the policy making process include promotion of transparency concerning the destination and use of public expenditure, improvement in the quality of policy making, which becomes more evidence based, increase in the collaboration across government bodies, as well as between government and citizens, increase the awareness of citizens on specific issues, as well as their information about government policies, and promotes accountability of public officials. Nevertheless, transparency does not directly imply accountability. "A government can be an open government, in the sense of being transparent, even if it does not embrace new technology. And a government can provide open data on politically neutral topics even as it remains deeply opaque and unaccountable." (Robinson & Yu, 2012).

Technologies, tools and methodologies. An interesting topic of research is the integration of open government data, participatory sensing and sentiment analysis, as well as visualization of real-time, high-quality, reusable open government data. Other avenues of research include the provision of quality, cost-effective, reliable preservation and access to the data, as well as the protection of property rights, privacy and security of sensible data (e.g. Charalabidis et al. 2018). Inspiring cases include: Open Government Initiative[34] carried out by the Obama Administration for promoting government transparency on a global scale; Data.gov:[35] platform which increases the ability of the public to easily find, download, and use datasets that are generated and held by the Federal Government. In the scope of Data.gov, US and India have developed an open source version called the Open Government Platform[36] (OGPL), which can be downloaded and evaluated by any national Government or state or local entity as a path toward making their data open and transparent; USAspending.gov:[37] it is a searchable website displaying for each Federal award the name of the entity receiving the award, the amount of the award, information on the award, and the location of the entity receiving the award; FederalRegister.gov:[38] HTML Edition of the Federal Register to make it easier for citizens and communities to understand and get informed about the regulatory process; performance.gov: [39] website providing a window of US Government Administration effort to improve performance and accountability.

[34] www.whitehouse.gov/open
[35] www.data.gov/
[36] www.opengovplatform.org/
[37] www.usaspending.gov/
[38] For more information please refer to www.federalregister.gov/
[39] For more information please refer to www.performance.gov/

Research perspectives. An extremely extensive taxonomy of Open Government Data research areas and topics is provided by Charalabidis et al. (2016), which distinguish 4 main research areas and 35 research topics. A high level representation of their taxonomy is depicted in Table 17.

Table 17 – Taxonomy of Open Government Data research areas and topics

OGD Management and Policies	OGD Infrastructures	OGD Interoperability	OGD Usage and Value
• Policy & Legal Issues for OGD • OGD Anonymisation Methods • OGD Cleaning Methods • OGD Quality Assessment Frameworks • OGD Visualisation methods and tools • OGD Linking • OGD Publishing • OGD Mining • OGD Rating and Feedback	• OGD Portals Architecture • Open Web Services / APIs • OGD User Profiling and Service personalisation • OGD Long-term Preservation • OGD Storage • Cloud computing for OGD • Citizen-generated open data • Sensor-generated open data	• Metadata for OGD • Multi-linguality • Service Interoperability Standards • Semantic Annotation • Ontologies • Platform technical Interoperability • Organisational Interoperability • Controlled Vocabularies and Code lists Preservation	• Skills Management for OGD • Reputation Management • OGD Use • OGD-based Entrepreneurship • OGD Value and Impact Assessment • OGD Needs Analysis • OGD-based Accountability • OGD Readiness Assessment • OGD Portals Evaluation Frameworks • OGD Innovation

Research Challenge 1.4 – Manipulation of statements and misinformation

Definition. Clearly transparency of policy making and overall trust can be negatively affected by fake news, disinformation and misinformation in general. In a more general sense disinformation can be defined as false information that is purposely spread to deceive people, while misinformation deals with false or misleading information (Lazer et al., 2018), but it also includes the bias that is inherent in news produced by humans with human biases. Lazer et al. (1094) define this most recent phenomenon as 'fabricated information that mimics news media content in form but not in organizational process or intent.' This is hardly a modern issue: what changes in the era of big data, is the velocity according to which fake news and false information spread through social media (e.g. Vaidhyanathan 2018). Another example related to big data technologies and that will become even more crucial in the future is the one of deepfakes (portmanteau of "deep learning" and "fake"), which is an artificial intelligence-based human image synthesis technique used to combine and superimpose existing images and videos onto source images or videos.

Relevance and applications in policy making. Fake news and misinformation lead to the erosion of trust in public institutions and traditional media sources, and in turn favour the electoral success of populist or anti-establishment parties. In fact, as discussed in Allcott and Gentzkow (2017) and Guess et al. (2018), Trump voters were more likely to be exposed and believe to misinformation. In the Italian context, il Sole 24 Ore[40] found that the consumption of fake news appear to be linked with populism,

[40] https://www.infodata.ilsole24ore.com/2018/05/04/fake-news-le-bugie-le-gambe-lunghe/

but the content of the overwhelming majority of pieces of misinformation also displays an obvious anti-establishment bias, as found in Giglietto et al. (2018). In the recent 2016 US presidential election, there has been the creation and spread of news articles that favoured or attacked one of the two main candidates, Hillary Clinton and Donald Trump, in order to steer the public opinion towards one candidate or the other. Furthermore, the success of Brexit referendum is another example of how fake news steered the public opinion towards beliefs that are hardly funded on evidence, e.g. the claim that UK was sending £350m a week to the EU, and that this money could be used to fund NHS instead.

Technologies, tools and methodologies. In the short term, raising awareness regarding fake news can be an important first step. For instance, the capability to judge is a source is reliable or the capability to triangulate different data sources is crucial in this regard. Furthermore, educating people on the capabilities of AI algorithms will be a good measure to prevent the bad uses of applications like FakeApp having widespread impact. For what concerns technologies for content verification, there are tools based on crowdsourced verification like CheckDesk, repositories of checked facts like FactCheck, and citizen journalism such as Citizen Desk. For what concerns verification platforms some very famous are SAM[41], Verily[42], Truly Media[43] and Check[44]. Plugins and browser tools include InVID[45] and Frame by Frame[46], Jeffrey's Image Metadata Viewer[47], Video Vault[48], NewsCheck[49] and RevEye[50]. Plugins that monitor social media and web content include the are Storyful's Multisearch[51] plug-in for searching Twitter, YouTube, Tumblr, Instagram and Spokeo, and Distill[52] which monitors web pages. On the other hand automated fact-checking tools include Chequado[53], ContentCheck[54], FullFact[55], Duke University's Reporters Lab[56], and Factmata[57]. Other very interesting examples are WeVerify[58], which is a blockchain database of known false claims and fake content, and Storyzy[59], which is a database of fake news sites and video channels.

Regarding technologies to counter fake news, NLP can help to classify text into fake and legitimate instances. In fact, NLP can be used for deception detection in text, and fake news articles can be considered as deceptive text (Chen et al., 2015; Feng et al., 2012; Pérez-Rosas and Mihalcea, 2015). More recently, deep learning has taken over in case large-scale training data is available. For what concerns text classification, feature-based models, recurrent neural networks (RNNs) models, convolutional neural networks (CNNs) models and attention models have been competing (Le and

[41] https://www.samdesk.io/
[42] https://veri.ly/
[43] http://www.truly.media/
[44] https://meedan.com/en/check/
[45] https://www.invid-project.eu/tools-and-services/invid-verification-plugin/
[46] https://chrome.google.com/webstore/detail/frame-by-frame-for-youtub/elkadbdicdciddfkdpmaolomehalghio
[47] http://exif.regex.info/exif.cgi
[48] https://www.bravenewtech.org/
[49] https://firstdraftnews.org/launching-new-chrome-extension-newscheck/
[50] https://chrome.google.com/webstore/detail/reveye-reverse-image-sear/keaaclcjhehbbapnphnmpiklalfhelgf
[51] https://chrome.google.com/webstore/detail/storyful-multisearch/hkglibabhninbjmaccpajiakojeacnaf
[52] https://chrome.google.com/webstore/detail/distill-web-monitor/inlikjemeeknofckkjolnjbpehgadgge
[53] https://chequeado.com/
[54] https://team.inria.fr/cedar/contentcheck/
[55] https://fullfact.org/
[56] https://reporterslab.org/
[57] https://factmata.com/
[58] https://twitter.com/WeV3rify/status/1044876853729796099
[59] http://storyzy.com/about

Mikolov, 2014; Zhang et al., 2015; Yang et al., 2016; Conneau et al., 2017; Medvedeva et al., 2017). Clearly all leading machine learning techniques for text classification, including feature-based and neural network models, are heavily data-driven, and therefore require quality training data based on sufficiently diverse and carefully labelled set of legitimate and fake news articles. Regarding deepfakes, another possibility is to make use of blockchain technologies, in which every record is replicated on multiple computers and tied to a pair of public and private encryption keys. In this way, the person/institution holding the private key will be the true owner of the data, not the computers storing it. Furthermore, blockchains are rarely affected by security threats, which in turn can attack centralized data stores. As an example, individuals could make of the blockchain to digitally sign and confirm the authenticity of a video or audio file. The more the digital signatures, the more is the likelihood that a document is authentic. Lately, DARPA has been heavily investing in algorithms for deepfake detection, as the subject is becoming more and more a matter of national security.[60]

Research perspectives. There are several research domains that are being explored in order to be able to contrast manipulation of statements and misinformation. From one side, there is strand of research on the effects of misinformation of memory, especially by mean of neuroimaging and other neuroscientific measurement techniques. On the other hand, there several strands of research devoted to counter manipulation of statements and misinformation. For instance, new forms of regulation are under scrutiny. Further, new methodologies for bot detection are being developed, such as a system developed by Fraunhofer that it automatically analyzes social media posts, deliberately filtering out fake news and disinformation, through the use of machine learning techniques and drawing on user interaction to optimize the results as it goes.[61] On the same line, Fabula has patented what it dubs a "new class" of machine learning algorithms to detect "fake news" in the emergent field of geometric deep learning, where the datasets to be studied are so large and complex that traditional machine learning techniques struggle to work, like in the case of patterns on complex, distributed data sets like social networks.[62] Finally, for what concerns deepfakes, another strand of research concerns making use of blockchain technologies, in which every record is replicated on multiple computers and tied to a pair of public and private encryption keys. In this way, the person/institution holding the private key will be the true owner of the data, not the computers storing it. Furthermore, blockchains are rarely affected by security threats, which in turn can attack centralized data stores. More synthetically, according to the Panel for the Future of Science and Technology of the European Parliamentary Research Service (2019), academic research is focused on three classes of approaches. The first set focuses on investigating the influence of social media platforms and online news sites and their influence in creating partisanship echo chambers, e.g. generating homogenous and polarised echo chambers (Del Vicario et al., 2016), or confirmation bias (Quattrociocchi, Scala & Sunstein, 2016). The second strand of research focused on detecting fake amplifiers of false narratives, through bots used in amplifying false narratives (Howard & Kollanyi, 2016; Gorrell et al, 2018). The third strand of work is on combining content analysis with network analysis through the use of semantic tools and machine learning (Conroy, Rubin & Chen, 2015) to assess veracity of information. Further, the Panel for the Future of Science and Technology distinguishes five areas:

[60] https://futurism.com/darpa-68-million-technology-deepfakes
[61] https://www.fkie.fraunhofer.de/en/press-releases/software-that-can-automatically-detect-fake-news.html
[62] https://techcrunch.com/2019/02/06/fabula-ai-is-using-social-spread-to-spot-fake-news/

- Fact checking and Content Verification: this deals with the use of automated fact-checking methods based on Natural Language Processing (NLP) and Artificial Intelligence (AI) techniques for fact checking and content verification;
- Detecting Computational Amplification and Fake Accounts: this deals with algorithms detecting bot and sockpuppet accounts, clickbait, and astroturfing;
- Detecting Mis- and Disinformation Campaigns: this deals with the research challenge of verifying not only the authenticity of online content, but also to find the originating source of viral disinformation, track its spread across online communities and networks, and establish the likely impact on citizens;
- Hate Speech, Abuse and Trolling: social platforms have started implementing semi-automated solutions to screen efficiently the large number of posts and comments received. For instance, Facebook have implemented machine learning algorithms that can identify abusive language;
- Accuracy and Effectiveness: social platforms and researchers are actively developing methods based on machine learning algorithms, in order to identify automatically disinformation on social media platforms. However, challenges are given by algorithmic scalability and possibility to make mistakes.

5.3.2 Research Challenges on Public Governance Framework for Data Driven Policy Making Structures

Research Challenge 2.1 - Forming and monitoring of societal and political will

Definition and applications in policy making. Many efforts have been undertaken by European governments to establish data platforms and of course, the present development in the open data movement contributes to data driven decisions in the public sector, but is the status quo sufficient or what is needed to leverage data for an advanced data based decision support in the public sector? The legislative and political objectives are often neither clear nor discussed in advance. This can lead to the point, that a huge amount of data is certainly available but not the right data sets to assess specific political problems. In that sense, governance structures and frameworks should be able to make the right data available and furthermore, should ensure that data analyses are interpreted bearing in mind societal and legislative goals and values (Schmeling et al.)

Further, data driven policy making is often discussed along with evidence based policy making. At its heart evidence based policy means that research results are applied by the policy making system (Wilsdon and Doubleday, 2015). On the other hand, data driven policy making is discussed in the sense of controlling and impact assessment through domain specific indicator systems (OECD, 2014).

Technologies, tools and methodologies.

Objectives in the public sector can be multifarious since they are aimed at the common good and not only prior at profit maximisation. Therefore, shared targets are a methodology that has the potential to transform common policies and legislative intentions on a horizontal and a vertical level into public organisations (James and Nakamura, 2015) Strategic instruments like the balanced score card or the canvas Model[63] are tools to support the operationalisation and transformation of targets into actions and indicators.

[63] http://thegovlab.org/introducing-the-digital-policy-model-canvas/

Innovative tools like experimentation, simulation, regulatory sandboxes or systems thinking can be applied for policy making in order to provide profound research insights for decision makers (OECD, 2018). An interesting case for policy simulations is given by the TNO policy lab for the co-creation of data-driven policy making. The Policy Lab is a methodology for conducting controlled experiments with new data sources and new technologies for creating data-driven policies. Policy makers experiment with new policies in a safe environment and then scale up. The Policy Lab approach has three pillars: (1) the use of new data sources as sensor data and technological developments for policy development; (2) a multidisciplinary approach: including data science, legal expertise, domain knowledge, etc.; and (3) involving citizens and other stakeholders ('co-creation') and carefully weighing different values (van Fleur Veenstra and Kotterink, 2017).

Research Perspectives.
Having in mind the two perspectives, the science and the controlling/monitoring perspective, research is needed in questions of how governance frameworks can balance between these both. On the one hand the need to install early warning and indicator monitoring as continuous evaluation systems and on the other hand the need to explain why thinks happen in example through randomised control trials, forecasts or simulations of political programmes.

Facing the high time pressure of the daily political life, it has to be investigated how and at what point to apply new data science methodologies based on machine learning and AI in order make controlled experiments and other methodological designs as efficient as possible. (Wilsdon *et al.*, 2015, p. 17) Furthermore, it is important to gather evidence more broadly from all sources – including from citizens – and then discuss more openly how to weigh and balance the different aspects. Even if science points in one direction, society might decide to go in another (Madelin, 2015, p. 29). Policy decisions should base on scientific evidence along with both society's value preferences and political judgment (Madelin, 2015, p. 27). It has to be investigated how political and societal will can be expressed and operationalized in order to be able to design monitoring systems and performance measurement systems based not simply on financial information but rather on outcome and performance-oriented indicators and research results.

Research Challenge 2.2 - Stakeholder/Data-producer-oriented Governance approaches
Definition. To enhance the evidence-based decisions in policy making, data must be gathered from different sources and stakeholders respectively including commercial data, citizens' data, third sector data and public administrations' data. Every Stakeholder group requires different approaches to provide and exchange data. These approaches must consider political, administrative, legal, societal, management and ICT related conditions. As a plurality of independent stakeholder groups is involved in the fragmented process of data collection, the governance mode cannot be based on a hierarchical structure. Thus, the network governance approach applies rather on negotiation-based interactions that are privileged to aggregate information, knowledge and assessments that can help qualifying political decisions (Sørensen and Torfing, 2007). The public administration is in its origin an important advisor of the political system and is not to be underestimated in this context, since the administration owns meaningful data, which should be considered profoundly in political decision making. In addition, the roles and responsibilities of public administrations as data providers must be discussed and clarified.[64]

[64] For a discussion of the role of government in data trading see Virkar et at. 2019

If specific company data like traffic data from navigation device providers or social media data from social network providers or one step further purchase data of medicine is necessary to assess political questions or early warning systems for public health, guidance, governance models and legal frameworks to purchase or exchange these data are needed (Micheli et al. 2018, Rinnerbauer et al. 2018). Moreover, for all aforementioned cases IT standards and IT architecture frameworks for processing data stored in different infrastructures constituting so called data spaces are required (Cuno et al., 2019). In this regard, an example is played by massive interconnections (massive number of objects/things/sensors/devices) through the information and communications infrastructure to provide value-added services, in particular in the context of smart cities initiatives. The unprecedented availability of data raises obvious concerns for data protection, but also stretches the applicability of traditional safeguards such as informed consent and anonymization (see Kokkinakos et al. 2016). Data gathered through sensors and other IoT typically are transparent to the user and therefore limit the possibility for informed consent, such as the all too familiar "accept" button in websites. Secondly, the sheer amount of data makes anonymization and pseudonymisation more difficult as most personal data can be easily deanonymized. Advanced techniques such as multiparty computation and homomorphic encryption remain too resource intensive for large scale deployment. We need robust, modular, scalable anonymization algorithms that guarantee anonymity by adapting to the input (additional datasets) and to the output (purpose of use). Additionally, it is important to ensure adequate forms of consent management across organization and symmetric transparency, allowing citizens to see how their data are being used, by whom and for what purpose. Clearly sometimes the options are limited, as in the case of geo-positioning, which is needed to be able to use the services provided. Basically, in this case the user pays with their data to use services.

Relevance and applications in policy making. Big data offer the potential for public administrations to obtain valuable insights from a large amount of data collected through various sources, and the IoT allows the integration of sensors, radio frequency identification, and Bluetooth in the real-world environment using highly networked services. The trend towards personalized services only increases the strategic importance of personal data, but simultaneously highlights the urgency of identifying workable solutions. On the other hand, when talking about once only principle, bureaucracy and intra-organisational interoperability are far more critical.

Technologies, tools and methodologies. Several tools are today being developed in this area and should be considered by public governance frameworks. Blockchain providing an authentication for machine to machine transaction: blockchain of things. More specifically, inadequate data security and trust of current IoT are seriously limiting its adoption. Blockchain, a distributed and tamper-resistant ledger, maintains consistent records of data at different locations, and has the potential to address the data security concern in IoT networks (Reyna et al. 2018). Anonymization algorithms and secure multiparty mining algorithm over distributed datasets allow guaranteeing anonymity even when additional datasets are analysed and the partitioning of data mining over different parties (Selva Rathna and Karthikeyan 2015).

Edge Computing as an additional technology has special relevance to the autonomous cyber-physical systems in the IoT environment, such as autonomous vehicles, implanted medical devices, fields of highly distributed sensors, and mobile devices. Cyber-physical systems need a huge amount of real-time data just to be usable. In an edge computing layer the real-time relevant data can be processed, filtered, stored and analysed locally and just relevant results can be sent to central data servers or cloud layers,

respectively. In this manner Edge Computing helps to relieve conventional computing systems and ensures lower latencies (International Electrotechnical Commission, 2017, p. 105). Therefore Edge Computing and especially the use of micro data centres should be also considered as potentially important in questions of data ownership and data governance.

Research Perspectives.
Yet, we do not have reached the maturity to share and add value to data among public sector, private sector and general public. A European Data Strategy and IT architecture reference models, like the Open Urban Platform (DIN SPEC 91357) have to be improved to establish alternative data governance models which includes collection, storing and sharing of data through multi-stakeholder data sharing agreements, commons-based data crowdsourcing or city data commons. (Micheli et al., 2018) Power relations have to be regulated between actors to develop structures that do not just pave the way to a surveillance capitalism where data is just a source of revenue. (Zuboff, 2015)
Also the instrument of incentives could be used to promote data sharing between individuals, companies, and public administrations. Furthermore success stories are required to showcase how data can be leveraged in different application scenarios, for different stakeholder groups and with different stories data is able to tell. (Micheli *et al.*, 2018)
The standardization work for urban and public data spaces at various relevant standardization bodies and relating to different domains has to be evolved. In addition, the reference implementation of standard open source components for European Data Spaces would allow quickly setting up (Cuno *et al.*, 2019).

Research Challenge 2.3 - Governance administrative levels and jurisdictional silos
Definition. Decisions in the political environment are often facing trans-boundary problems on different administrative levels and in different jurisdictions (Micheli *et al.*, 2018, p. 2). Thus, the data collection to understand these problems and to investigate possible solutions causes manifold barriers and constraints, which have to be overcome through modern governance approaches and models. Like the aforementioned stakeholder network of data providers, a data network has to be coordinated on meta-level and respective rules and access rights have to be established ICT-enabled through data connectors or controlled harvesting methods. This is becoming increasingly urgent as government holds massive and fastly growing amounts of data that are dramatically underexploited. The achievement of the once only principle, as well the opportunities of big data only add to the urgency. Interoperability of government data, as well as the issues of data centralization versus federation, as well as data protection, remain challenges to be dealt with.
Relevance and applications in policy making. Data integration has long been a priority for public administration but with the new European Interoperability Framework and the objective of the once only principle is has become an unavoidable priority. Data integration and integrity are the basic building blocks for ensuring sufficient data quality for decision-makers – when dealing with strategic policy decision and when dealing with day to day decisions in case management.

Technologies, tools and methodologies.
New interfaces within which the single administrations can communicate and share data and APIs in a free and open way, allowing for the creation of new and previously-unthinkable services and data applications realised on the basis of the needs of the citizen. **C**losely integrated services across agencies increase their use as demonstrated in Estonia's X-Road framework which integrates services from all

parts of government as well as established protocols that govern data exchange or security standards (World Bank, 2016, p. 35) The X-Road[65] is an infrastructure which allows the Estonian various public and private sector e-service information systems to link up. Currently, the infrastructure is implemented also in Finland, Kyrgyzstan, Namibia, Faroe Islands, Iceland, and Ukraine. Another interesting example is given by 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.[66]

Research Perspectives.
Technology interacts with rules such as regulations and standards which are established locally (World Bank, 2016, p. 28). So technological frameworks must be developed to become flexible enough to adapt to these different regulations and standards.
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 autonomy, and the requirement for ex ante homogenization with more pragmatic, on demand approaches based on the "data lake" paradigm. All this need to take place at European level, to ensure the achievement of the goals of the Tallinn declaration. And appropriate, modular data access and interoperability is further complicated by the need to include private data sources as provider and user of government data, at the appropriate level of granularity. Last but not least, this needs to work with full transparency and full consent by citizens, ideally enabling citizens to track in real time who is accessing their personal data and for what purposes.

Research Challenge 2.4 - Education and personnel development in data sciences
Definition. Governance plays also an important role on all questions of education and personnel development in order to ensure that the right capabilities and skills are available in terms of data literacy, data management and interpretation (Lewis and Pettersson, 2009). The need to develop these skills has to be managed and governed as a basis to design HR strategies, trainings and employee developments. It is also to ensure that infrastructure investments get completely be leveraged and to do so accurate, relevant and representative data is required (Chetty *et al.*, 2018, p. 3). Considering that Just 19 of 1000 individuals are graduated in STEM (Science, Technology Engineering and Mathematics) disciplines in Europe and gap between demand and supply of ICT specialists in the EU is expected to widen further, the need to improve educational programs in this field to ensure lifelong learning becomes obvious (European Commission, 2018, p. 3)

Relevance and applications in policy making. Governance in personnel development promotes effective and efficient fulfilment of public duties like evidence based policymaking. This is all the more true when taking into account the use of Big Data in policy making, as clearly the skills and competence of civil servants and politicians are very important for the implementation of reforms and take up of data strategies and solutions (Lewis and Pettersson, 2009). For policymakers and regulators it is important to become clear about what policy tools could be used to incentivize digital initiatives that deliver value to society and whether the right digital skills and talents are in place and in this context how to exchange lessons learnt from the experience of private-sector organisations. (World Economic Forum, 2016)

[65] https://www.bigpolicycanvas.eu/community/kb/x-road
[66] https://www.bigpolicycanvas.eu/community/kb/italian-data-analytics-framework-daf

Technologies, tools and methodologies. This research challenge includes focusing on standards to make transparent the assessment criteria of education policies, incentives to motivate specific types of behaviour, information in the way of clear definitions of outputs and outcomes and accountability to examine that given outcomes and outputs can be delivered (Lewis and Pettersson, 2009).

Research Perspectives.
The Open Data Barometer revealed that about 85 percent of developing countries had made little or no progress in opening map data. Reasons encompass lack of technical skills, inadequate resources, and unwillingness to expose data to scrutiny. (World Bank, 2016, p. 28) Education can help to de-mystify technology and data analysis. There is a pressing need also for ethical education, both for developers and designers, and for the public at large. Arguably, this is even more crucial when we have machines that can make complicated and consequential decisions. (Micheli *et al.*, 2018, p. 25) To ensure data literacy and digital literacy both should be monitored by governments through a digital literacy index which should be developed and elaborated further (Chetty *et al.*, 2018) Therefor multidisciplinary data collection instruments should be developed and included in respected governance structures (Chetty *et al.*, 2018, p. 12).
The digital environment also requires an analogue foundation, in the form of regulations that create a climate which allows public servants to seize their skills in the digital world and public institutions to use the internet to empower citizens. (World Bank, 2016, p. 5) It has to be more investigated how technology interacts with other factors like the necessity of human judgement, intuition or discretion. Technology interacts with workers' skills. Only if the right routine tasks get automated, workers can leverage technology to become more productive by focusing on personal interaction, scheduling and other tasks. (World Bank, 2016, p. 28)

5.3.3 Research Challenges on Data acquisition, Cleaning and Representativeness

Research Challenge 3.1 - Real time big data collection and production
Definition. The rapid development of the Internet and web technologies allows ordinary users to generate vast amounts of data about their daily lives. On the Internet of Things (IoT), the number of connected devices has grown exponentially; each of these produces real-time or near real-time streaming data about our physical world. In the IoT paradigm, an enormous amount of networking sensors are embedded into various devices and machines in the real world. Such sensors deployed in different fields may collect various kinds of data, such as environmental data, geographical data, astronomical data, and logistic data. Mobile equipment, transportation facilities, public facilities, and home appliances could all be data acquisition equipment in IoT. Furthermore, social media analytics deals with collecting data from social media websites like Facebook, Twitter, YouTube, WhatsApp etc. and blogs. Social media analytics can be categorized under big data because the data generated out of the social websites are in huge number, so that some efficient tools and algorithms are required for analysing the data. Data collected include user-generated content (tweets, posts, photos, videos), digital footprints (IP address, preferences, cookies), Mobility data (GPS data), Biometric information (fingerprints, fitness trackers data), and consumption behaviour (credit cards, supermarket fidelity cards) (see for example, (Xu, Wang, Peng, & Wu, 2019).

Relevance and applications in policy making. The collection of such amounts of data in real time can help in updated evaluation of policies, in monitoring the effects of policy implementations, in collecting data that can be used for agenda setting (for instance traffic data), as well as for the analysis of the sentiment and behaviour of the citizens, monitoring and evaluating government social media communication and engagement. Specifically, IoT has extensive applications in the smart city realm, for instance in:

- Smart car parking: sensors can detect availability and signal to the network that processes the vacancy map to provide it to the parking manager or even directly to the driver;
-
- Waste management: sensors integrated in the waste containers are able to detect the filling level to optimize the collection paths;
- Road viabilities: monitoring of the flow of vehicles and the number of pedestrians for the collection and processing of data to improve driving and pedestrian routes;
- Water management: both in terms of monitoring the level of water pollution and in terms of managing water flows to prevent floods.
- Intelligent lighting: lamps equipped with sensors, light up only when passing cars, pedestrians or bikes. They also recognize atmospheric conditions to ensure the ideal degree of illumination;
- Structural control: monitoring of vibrations and material conditions in buildings, bridges and historical monuments;
- Noise maps: monitoring sound pollution in to build noise maps;
- Air quality: measurement much more efficient and spread out in the territory;
- Smart roads: intelligent highways with warning messages and deviations based on weather conditions or in the event of unforeseen events such as accidents or traffic jams;
- Maintenance: all kinds of equipment (means of transport, lamps, electronic devices) can communicate remotely with those who manage it, communicating their operating status, thus allowing intelligent maintenance plans to be defined.

On the other hand, for what concerns the use of social media data, an obvious element is the analysis of government use of social media aimed to monitoring and evaluating government social media communication and engagement. Another interesting element, would be the analysis and research of the public's use of social media, helping to support the development, implementation, and evaluation of government policy and service delivery.

Technologies, tools and methodologies. For collecting the data from devices, as already mentioned Internet of Things technologies can be integrated in smart city platforms. A very interesting example is the European-funded IoT open source platform FIWARE[67], which is an open source initiative defining a universal set of standards for context data management which facilitate the development of Smart Solutions for different domains such as Smart Cities, Smart Industry, Smart Agrifood, and Smart Energy. Specifically, the FIWARE IoT platform provides a set of APIs and also combines components enabling the connection to the Internet of Things with Context Information Management and Big Data services on the Cloud. Another (proprietary) platform is the CISCO Kinetic for Cities[68], which is an end-to-end data platform that consists of three different offerings: connectivity solutions enable users to

[67] https://www.fiware.org
[68] https://www.cisco.com/c/en/us/solutions/industries/smart-connected-communities/kinetic-for-cities.html

gather data from sensors and IoT endpoints; the IoT platform transforms, ingests, normalizes, and transports data to applications and users; and the security portfolio provides access to a city's network and data platform.

Regarding social media, there are many collection and analytics tools readily available for collecting and analysing content. These tools help in collecting the data from the social websites and its service not only stop with data collection but also helps in analysing the usage of data. Examples of tools and technologies are online sentiment analysis and data mining, APIs, data crawling, data scraping What is interesting about the development of such tools, is the development of automated technological tools that can collect, clean, store and analyse large volumes of data at high velocity. Indeed, in some instances, social media has the potential to generate population level data in near real-time. Methodologies used to produce analysis from social media data include Regression Modelling, GIS, Correlation and ANOVA, Network Analysis, Semantic Analysis, Pseudo-Experiments, and Ethnographic Observations. In the same way, following the Social Media Research Group (2016) and Veltri (2019), tools for text analysis include:

- DiscoverText: cloud-based text analysis tool allows different data (including, but not limited to twitter data) to be stored in different project folders;
- NCapture: this is an NVivo 10 add-on allowing users to capture data from their web browser, such as segmented Twitter data, for analysis;
- SentiStrength: this is a program that compares social media text against a lexicon-based classifier of sentiments, and provides a separate score for each word within a sentence thereby giving the average sentiment strength of the content.

For what concerns network analysis:

- Gephi: this is an open source program allowing visualisation and exploration of networks, including social media networks;
- Node XL: this is basically an Excel add-on which enables interactive exploration of network graphs, and that can be easily applied also to social media;
- SocSciBot: this program can be used to run limited analyses of the text in the websites aiming to produce statistics and diagrams explaining the interlinking of pages on websites;

For what concerns data acquisition tools:

- Tweet Archivist: this program creates archives of tweets, which can then be downloaded and visualized;
- ScraperWiki: this is a program that can be used to analyze and visualize a wide range of data from a hige number of web-based sources;

Finally, concerning multipurpose platforms:

- Pulsar: this is a platform performing data analysis from social media, such as time series, topic and sentiment analysis, and which provides also interesting visualisations;
  YouGov SoMA: this platform performs an overlay of demographic data with comments made on social media platforms such as Twitter and Facebook, to identify what an audience is paying attention to;
- Crimson Hexagon: this is a platform used for qualitative analysis, such as sentiment and text analysis and user backgrounds.

Research perspectives. Big data in terms of data collection by means of sensors and their combination with other sources appears to be the next evolution in the way the dimensionality of datasets available to researcher is massively increasing. The combination of 'sensors big data' with behavioural\usage ones and self-reported (e.g. user-generated content) is a unique possibility of tracing dynamic unfolding at different levels of measurements. It allows researcher a much finer view of processes, and that is a

new frontier to develop models that are better representations of complex adaptive systems, like the social and economic systems are.

Research Challenge 3.2 - Quality assessment, data cleaning and formatting

Definition. Big Data Quality assessment is an important phase integrated within data pre-processing. It is a phase where the data is prepared following the user or application requirements. When the data is well defined with a schema, or in a tabular format, its quality evaluation becomes easier as the data description will help mapping the attributes to quality dimensions and set the quality requirements as baseline to assess the quality metrics. After the assessment of data quality, it is time for data cleaning. This is the process of correcting (or removing) corrupt or inaccurate records from a record set, table, or database. Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant, etc. parts of the data and then replacing, modifying, or deleting this dirty data or coarse data (Lagoze, 2014) This research challenge also deals with formatting, as once one has downloaded sets of data is not obvious at all that their format will be suitable for further analysis and integration in the existing platforms. And another important factor is metadata, which are important for transparency and completeness of information.

Relevance and applications in policy making. Apart from systematic errors in data collection, it is important to assess to extent to which the data are of quality, and to amend it, obviously because policy decisions have to be funded on quality data and therefore have to be reliable. More data does not necessarily mean good or better data, and many of the data available lack the quality required for its safe use in many applications, especially when we are talking about data coming from social networks and internet of things. Apart from that, data to be used in analysis related to policy making and policy modelling should be duly formatted.

Technologies, tools and methodologies. Regarding data quality, it is mandatory to use existing and develop new frameworks including big data quality dimensions, quality characteristics, and quality indexes. For what concerns data cleaning, the need for overcoming the hurdle is driving development of technologies that can automate data cleansing processes to help accelerate business analytics. Considering frameworks for data quality assessment, the UNECE Big Data Quality Task Team released in 2014 a framework for the Framework for the Quality of Big Data within the scope of the UNECE/HLG project "The Role of Big Data in the Modernisation of Statistical Production" (UNECE 2014). Further, the quality framework provides a structured view of quality at three phases of the business process:

- Input, i.e. when the data is acquired, or in the process of being acquired (collect stage);
- Throughput, i.e. any point in the business process in which data is transformed, analysed or manipulated. This might also be referred to as 'process quality' (process and analyse stages);
- Output, i.e. the assessment and reporting of quality with statistical outputs derived from big data sources (evaluate and disseminate stage).

The framework is organized according to a hierarchical structure composed of three hyper-dimensions;

- Source: relates to factors associated with the type of data, the characteristics of the entity from which the data is obtained, and the governance under which it is administered and regulated.

This hyper-dimension includes quality dimensions such as institutional and business environment, and privacy and security;

- Metadata, which relates to the characteristics of the entity from which the data is obtained, the governance under which it is administered and regulated, as well as factors associated with the type of data. This hyper-dimension includes quality dimensions such as complexity, completeness, usability, time-related factors, linkability, coherence and consistency, and validity;
- Data, which relates to the quality of the data itself, includes quality dimensions such as accuracy and selectivity, linkability, coherence and consistency, and validity.

For what concerns techniques for data cleaning, there are several actions to be carried out:

- Setting up a data quality plan, defining clear KPIs along with areas where the data errors are more likely to occur and at the same time identifying the reasons for errors in the data;
- Remove duplicate observations, coming from the combination of datasets from multiple places or from scraping, as well as irrelevant observations, which are those that do not actually fit the specific problem at hand;
- Fix structural errors, such as typos or inconsistent capitalization;
- Eliminate unwanted outliers, which are basically data entries that affect the robustness of your model;
- Handle missing data, either by dropping observations that have missing values, or by imputing the missing values based on other observations.

Clearly all available software have routines for performing data cleaning, even though the process requires always a direct observation of the dataset.

Research perspectives. Concerns about digital data and their quality are understandable. The impression is that in current digital methods there is an unbalance between the attention given to data collection and analysis and that paid to issues of errors, validity and general quality. Perhaps it is a normal sequence of development: first, new ways of getting and analysing data emerge; then, concerns about their validity and robustness become more salient. Any effort to develop quality standards for big data research is something that is, at the moment, lacking. Therefore, quality standards are where the next effort should be allocated. At the same time, techniques to assess data quality need to be innovated given the complexity of the dataset now available. Development in topological analysis, for example, are very promising (Snášel, Nowaková, Xhafa, & Barolli, 2017).

Research Challenge 3.3 - Representativeness of data collected

Definition. A key concern with many Big Data sources is the selectivity, (or conversely, the representativeness) of the dataset. A dataset that is highly unrepresentative may nonetheless be useable for some purposes but inadequate for others. Related to this issue is the whether there exists the ability to calibrate the dataset or perform external validity checks using reference datasets. Selectivity indicators developed for survey data can usually be used to measure how the information available on the Big Data Source differs from the information for the in-scope population (Braun & Kuljanin, 2015). For example, we can compare how in-scope units included in Big Data differ from in-scope units missing from the Big Data. To assess the difference, it is useful to consider the use of covariates, or variables that contain information that allows to determine the "profile" of the units (for example, geographic location, size, age, etc.) to create domains of interest. It is within these domains that comparisons should be made for "outcome" or study variables of interest (for example, energy consumption, hours worked, etc.). Note

that the covariates chosen to create the domains should be related to the study variables being compared. Regarding social media, research defines a set of challenges that have implications for have implications for validity and reliability of data collected. First, users of social media are not representative of populations (Ruths & Jurgen, 2014). As such, biases will exist, and it may be difficult to infer findings to the general population. Furthermore, social media data is seldom created for research purposes, and finally it is difficult to infer how reflective a user's online behaviour is of their offline behaviour without information on them from other sources (Social Media Research Group 2016).

Relevance and applications in policy making. Clearly big data representativeness is crucial to policy making, especially when studying certain characteristics of the population and in analysing its sentiment (Salganik, 2019). It is also important of course when tackling certain subgroups. In this regard, large datasets may not represent the underlying population of interest and sheer largeness of a dataset clearly does not imply that population parameters can be estimated without bias.

Technologies, tools and methodologies. Appropriate sampling design has to be applied in order to ensure representativeness of data and limit the original bias when present. Probability sampling methodologies include: simple random sampling, stratified sampling, cluster sampling, multistage sampling, and systematic sampling. An interesting research area is survey data integration, which aims to combine information from two independent surveys from the same target population. Kim et al. (2016) propose a new method of survey data integration using fractional imputation, and Park et al. (2017) use a measurement error model to combine information from two independent surveys. Further, Kim and Wang (2018) propose two methods of reducing the selection bias associated with the big data sample. Finally, Tufekci (2014) provides a set of practical steps aimed at mitigating the issue of representativeness, including: targeting non-social dependent variables, establishment of baseline panels to study people's behaviour, use of multidisciplinary teams and multimethod/multiplatform analysis. Big Data can be also combined with 'traditional' datasets to improve representativeness (Vaitla 2014).

Research perspectives. The relationship between big data and representativeness is also relevant in a context in which methods are developed by the private sector, where business entities invest in data collection and analysis with more resources than many universities combined. The inequality is generated not only by different opportunities for accessing data, but also by the availability of research infrastructures, for example computing power, that is needed to fully exploit digital and big data. Even worse, academics in developing countries are even less likely to have access to digital data and opportunities for them to do research about their own societies are limited. It is not by chance that the large majority of studies about the so-called Arab Spring and the role of social media platforms have been carried out by Western academics. There is the need to avoid the temptation of doing digital research only on WEIRD (Western, Educated, Industrialized, Rich, and Democratic) participants or by WEIRD academics on anybody else (Henrich, Heine, & al, 2010). It is foreseeable that new arrangements will be necessary to strike a balance between data as assets for private and public institutions and, at the same time, access to the largest pool possible of researchers. The pressure to find a balance between the open access model and the protection of intellectual property will be particularly intense in the near future.

5.3.4 Research Challenges on data storage, clustering, and integration

Research Challenge 4.1 – Big Data Storage

Definition. A pre-requisite for clustering and integration of big data is the presence of efficient mechanisms for data storage and processing. Big data storage technologies are a crucial enabler for advanced analytics that has the potential to transform society and the way critical decisions are made, also in terms of policies. One of the first things organizations have to manage when dealing with big data, is where and how this data will be stored once it is acquired. The traditional methods of structured data storage and retrieval include relational databases and data warehouses.

Relevance and applications in policy-making. The data acquired by the public administration, to be further used for analytics, modeling, and visualization, need to be stored efficiently and safely. In this regard, it is essential to understand the encryption and migration needs, the privacy requirements, as well as the procedures for backup or disaster recovery. Furthermore, big data storage and processing technologies can produce information that can enhance different public services

Technologies, tools, and methodologies. This research topic has been developing rapidly, delivering new types of massive data storage and processing products e.g., NoSQL knowledge bases. Based on the advances of cloud computing, the technology market is very developed in this area (for an overview, see Sharma, 2016). Crowdsourcing also plays an important role, and in the light of climate change and environmental issues, energy-efficient data storage methods are also a critical research priority (Strohbach et al. 2016). Furthermore, to automate complex tasks and make them scalable, hybrid human-algorithmic data curation approaches have to be further developed (Freitas and Curry 2016). More specifically, the most important technologies are: distributed File Systems such as the Hadoop File System (HDFS), NoSQL and NewSQL databases, and Big Data Querying Platforms. On the other hand, interesting tools are: Cassandra, Hbase (George, 2011), MongoDB, CouchDB, Voldemort, DynamoDB, and Redis.

Research Perspectives. One of the big challenges in the area of data storage is to protect data from counterfeiting, i.e., to make sure that data records are not forged to serve fraudulent or criminal purposes. This is particularly important when considering sensible and personal data or general data about specific events (e.g., signature of contracts, agreements, historical events, etc.). A possible solution could come from block-chain technologies applied (to ensure immutability of data records) to the general domain of knowledge production and transfer and information spreading: from casual conversations to brainstorming, from lectures to workshops, from the collective creation of documents or artworks to the organization of the cultural, historical and artistic heritage. Whenever information and knowledge are produced and transferred, it would be desirable to track the full processes to generate trust and fairness in the acknowledgment of the credits. Current blockchain technologies seem well adapted to face this challenge while being more scalable and environmental-friendly. Important challenges include: the interoperability of different blockchain-based platforms, the interplay between public and private blockchains, the question of the safe and decentralized storage of the information (for instance in a IPFS

(Inter Planetary File System), the questions related to the validation of identities and time-stamps, the regulations needed to infuse trust and reliability to future blockchain-based systems.

Research Challenge 4.2 - Identification of patterns, trends and relevant observables

Definition. This research challenge deals with technologies and methodologies**,** allowing businesses and policy-makers to identify patterns and trends of data both structured and unstructured that may not have been previously visible.

Relevance and applications in policy-making. The possibility to extract patterns and trends in data can help the policy-maker in having deep insight or discovering critical issues to be taken into account when developing the policy agenda. An interesting application is, for instance, anomaly detection, commonly used in fraud detection. For example, anomaly detection can identify suspicious activity in a database and trigger a response. There is usually some level of machine learning involved in this case.

Technologies, tools, and methodologies. One of the most used Big Data methodologies for identification of pattern and trends is data mining. Combination of database management, statistics, and machine learning methods useful for extracting patterns from large datasets. Some examples include mining human resources data to assess some employee characteristics or consumer bundle analysis to model the behavior of customers. It also has to be taken into account that most of the Big Data is not structured and have a massive quantity of text. In this regard, text mining is another technique that can be adopted to identify trends and patterns.

Research Perspectives The availability of huge amounts of data does not represent per se a general solution. The point is that data (also big data) tell us something about the past and the knowledge of the past is not always helpful in designing the future. Looking at the future with the eyes of the past could be misleading also for machines. Despite the recent dramatic boost of inference methods, they still crucially rely on the exploitation of prior knowledge and the problem of how those systems could handle unanticipated knowledge remains a great challenge. In addition, also with the present available architectures (feed-forward and recurrent networks, topological maps, etc.) it is difficult to go much further than a black-box approach and the understanding of the extraordinary effectiveness of these tools is far from being elucidated. Given the above-mentioned context it is important to make steps towards a deeper insight about the emergence of the new and its regularities. This implies: (i) conceiving better modelling schemes, possibly data-driven, to better grasp the complexity of the challenges in front of us; (ii) Wisely blending modelling schemes, inference methods and data into new platforms aimed at fostering a renewed dialogue between scientific research and policy making.

Research Challenge 4.3 - Extraction of relevant information and feature extraction

Definition. Summarizing data and meaning extraction to provide a near real-time analysis of the data. Some analyses require that data must be structured homogeneously before using them. Unlike humans, algorithms are not able to grasp nuances. Furthermore, most computer systems work better if multiple items are stored in identical size and structure. An efficient representation, access, and analysis of semi-structured data are necessary because, as a less structured design is more useful for specific analysis and purposes. Even after cleaning and error correction in the database, some errors and incompleteness will remain, challenging the precision of the study. On top of that gathering *better data* could be a better

strategy with respect to *more data.* Acquiring more data does not always allow for a deeper understanding. Better data can be gathered through suitable campaigns (also implementing gaming schemes) to engage citizens, enhance their awareness and trigger a bidirectional dialogue between citizens and institutions;

Relevance and applications in policy-making. While information and feature extraction could appear far from the policy process, it is a fundamental requirement to ensure the veracity of the information obtained and to reduce the effort from the following phases, providing the most comprehensive reuse of the data for purposes different from the one it was initially gathered. The data have to be adapted according to the use and analysis to be performed and the accompanying visualizations.

Technologies, tools, and methodologies. Bayesian techniques for meaning extraction; extraction and integration of knowledge from massive, complex, multi-modal, or dynamic data; data mining; scalable machine learning; principal component analysis. Tools include NoSQL, Hadoop, deep learning, rapidminer, keymine, R, python, and sensor data processing (fog and edge computing).

Research Perspectives Strategic thinking and decision-making are two very challenging tasks for cognitive abilities of human beings. To this end it is crucial to synthesize complex situations (as mirrored by the available data) so to extract relevant features and indications to steer the decision-making process. The problem can be phrased as the exploration of the open-ended and expanding conceptual space embedding the problem at hand to identify suitable solutions. Two main research avenues are worth exploring: (i) developing platforms for "Interactive Dialogues", exploiting AI and language technologies, where humans will be allowed to collectively search for progressively better solutions, also formulating queries engaging AI systems in turn-taking process; (ii) developing suitable modelling schemes of complex problems (e.g., urban mobility, middle income traps, developmental challenges, etc.), based on available data and historical records, able to let stakeholders to conceive and explore new scenarios. Applications can be foreseen in many different environments (technological progress, corporate strategies, decision-making, etc.).

5.3.5 Research Challenges on Modelling and Analysis with Big Data

Research Challenge 5.1 – Identification, acceptance and validation of suitable modelling schemes inferred from existing data

Definition. The traditional way of modelling started with a hypothesis about how a system acts. Then collect data to test the model. Traditionally, the amount of data collected was small since it rarely already existed, had to be generated with surveys, or perhaps imputed through analogies. Finally, statistical methods established enough causality to arrive at enough truth to represent the system. So deductive models are forward running, so they end up representing a system not observed before. On the other hand, with the current huge availability of data, it is possible to identify and create new suitable modelling schemes that build on existing data. These are inductive models that start by observing a system already in place and one that is putting out data as a by-product of its operation. For a simulation model, the modeler needs to understand the processes that are taking places, he/she needs to gather the relevant data, and then to observe the system at work. On the other hand, for instance in a machine learning model, the modeler needs data to start with. In this respect very often data scientists are thrown

at with data and the instruction to find something, and limited situational awareness is required by the modeller. An example of data modelling approach is given by data mining, as reported by Kim at al. (2017): using data-mining techniques, it is possible not only analyze a pattern or property of data in one dimension, but also identify a distribution function of the data from the pattern. Then, after validating the distribution function with data in the real world, users can acquire a ‘‘random number. generation’’ model that can be utilized in the process of prediction of future data patterns. An example of this mechanism is provided in Figure .

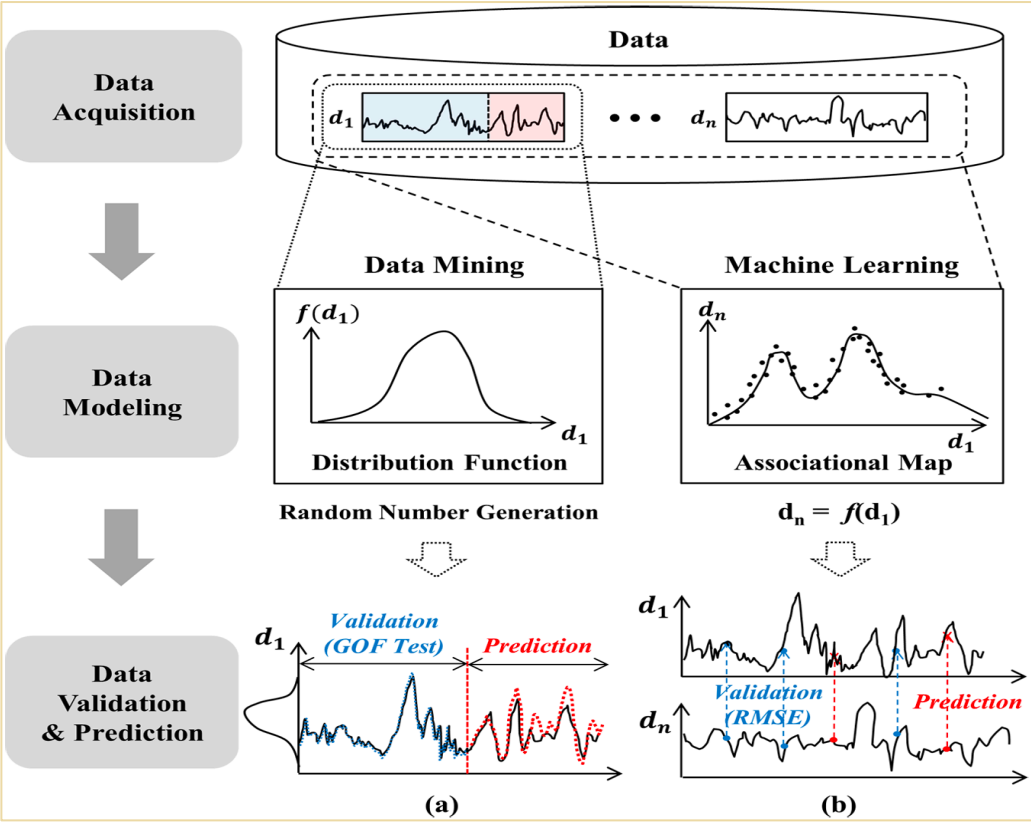

Figure 19 – Data Modelling Approach with Machine Learning

In this respect, the real challenge is to be able to identify, accept and validate from existing data models that are valid and suitable to cope with complexity and unanticipated knowledge. In fact, according to Mcintosh et al. (2008) there is a different perceptions of model users and model developers on what a model should look like. Furthermore, Van Delden et al. (2011) argue that model acceptance and validation is hindered by a lack of transparency, inflexibility and a focus on technical capabilities. Even the concept of model acceptance is not clear. In that regard McIntosh et al (2011) distinguish four levels of acceptance: model development has been completed and presented to its intended users; the users have been trained in the use of a model, but there is limited evidence of actual use; the model has been used on a one-off basis, and the model is used routinely in the daily work of the user.

Model acceptance and validation is composed of two main phases. The first phase is conceptual model validation, i.e. determining that theories and assumptions underlying the conceptual model are correct. A second phase is the computerised model verification, that ensures that computer programming and implementation of the conceptual model are correct. Specific to model validation, we can distinguish among conceptual validity, logical validity, experimental validity, operational validity, behavioral validity, representation validity and data validity.

Relevance and applications in policy making. There are several aspects related to the identification, acceptance and validation of modelling schemes that are important in policy making. A first deals with the reliability of models: policy makers use simulation results to develop effective policies that have an important impact on citizens, public administration and other stakeholders. Identification and validation is fundamental to guarantee that the output of analysis for policy makers is reliable. Another aspect is the acceleration of the policy modelling process: policy models must be developed in a timely manner and at minimum cost in order to efficiently and effectively support policy makers. Model identification and validation is both cost and time consuming and if automated and accelerated can lead to a general acceleration of the policy modelling process.

Technologies, tools and methodologies. In current practice the most frequently used is a decision of the development team based on the results of the various tests and evaluations conducted as part of the model development process. Another approach is to engage users in the choice and validation process. At any rate, conducting model validation concurrently with the development of the simulation model enables the model development team to receive inputs earlier on each stage of model development. Therefore, ICT Tools for speeding up, automating and integrating model validation process into policy model development process are necessary to guarantee the validity of models with an effective use of resources. It has finally to be noticed that model validation is not a discrete step in the simulation process. It needs to be applied continuously from the formulation of the problem to the implementation of the study findings as a completely validated and verified model does not exist. An example of validation procedures for simulation models is provided in Table 18.

Table 18 - Validation Procedures for Simulation Models

Procedure	Description
Face Validity	Consists of getting feedback from knowledgeable individuals about the phenomenon of interest through reviews, interviews, or surveys, to evaluate whether the (conceptual) simulation model and its results (input-output relationships) are reasonable.
Comparison to Reference Behaviors	Compares the simulation output results against trends or expected results often reported in the technical literature. It is likely used when no comparable data is available.
Comparison to Other Models	Compares the results (outputs) of the simulation model being validated to results of other valid (simulation or analytic) model. Controlled experiments can be used to arrange such comparisons
Event Validity	Compares the "events" of occurrences of the simulation model to those of the real phenomenon to determine if they are similar. This technique is applicable for event-driven models.
Historical Data Validation	If historical data exist, part of the data is used to build the model and the remaining data are used to compare the model behavior and the actual phenomenon. Such testing is conducted by driving the simulation model with either sample from distributions or traces, and it is likely used for measuring model accuracy.
Rationalism	Uses logic deductions from model assumptions to develop the correct (valid) model, by assuming that everyone knows whether the clearly stated underlying assumptions are true.
Predictive Validation	Uses the model to forecast the phenomenon's behavior, and then compares the phenomenon's behavior to the model's forecast to determine if they are the same. The phenomenon's data may come from the real phenomenon observation or be obtained by conducting experiments, e.g., field-tests for provoking its occurrence. Also, data from the technical literature may be used, when there is no complete data in hands. It is likely used to measure model accuracy
Internal Validity	Several runs of a stochastic model are performed to determine the amount of (internal) stochastic variability. A large amount of variability (lack of consistency)

	may cause the model's results to be questionable, even if typical of the problem under investigation.
Sensitivity Analysis	Consists of changing the values of the input and internal parameters of a model to determine the effect upon the model output. The same relationships should occur in the model as in the real phenomenon. This technique can be used qualitatively—trends only — and quantitatively —both directions and (precise) magnitudes of outputs.
Testing structure and model behavior	Submits the simulation model to tests cases, evaluating its responses and traces. Both model structure and outputs should be reasonable for any combination of values of model inputs, including extreme and unlikely ones. Besides, the degeneracy of the model's behavior can be tested by appropriate selection of values of parameters.
Based on empirical evidence	Collects evidence from the technical literature (experimental studies reports) to develop the model's causal relationships (mechanisms).
Based on empirical evidence	Individuals knowledgeable about the phenomenon are asked if they can distinguish between real and model outputs.

Regarding current applications concerning model acceptance and validation, Kolkman et al. (2016) identify a set of criteria within the scope of policymaking: model characteristics, organizational characteristics and supporting infrastructure. Model characteristics include: quality, which is the degree to which the model is perceived to be valid; the intuitive tractability of the chosen modelling technique; efficiency, meaning the ability to produce the model outcomes in a timely fashion; and flexibility, recognised as the potential ease with which a model can be adapted to inform new questions. Organizational factors include the capability to implement the model in a timely fashion, as well as the presence within an organization of advocates spearheading the model. And supporting infrastructure deals with the degree to which a model is implemented in a programming language or software platform that the user is familiar with, the transparency of the model in terms of ability to review the mechanics of the model and its underlying assumptions, and the consistency of the model results with similar existing models. Finally, they discuss two other very important criteria, namely the fact that the modeller has a strong reputation, and participation in development. This last one is very important, as clearly the opportunity to be involved in the process of developing the model, i.e. defining the mechanisms of a model on a conceptual level, contributes to the acceptance of the model.

Research perspectives. Regarding research perspectives for model choice, acceptance and validation, a clear aim is to develop ICT Tools for speeding up, automating and integrating model validation process into policy model development process in order to guarantee the validity of models with an effective use of resources. In fact, in order to speed up and reduce the cost of a model validation process, user-friendly and collaborative statistical software should be developed, possibly combined with expert systems and artificial intelligence. Further, given the big gap between theory and practice, the considerable opportunity exists for the study and application of rigorous verification and validation techniques. Complicated simulation models are usually either not validated at all or are only subjectively validated. Therefore, complexity issues in model validation may be better addressed through the development of more suitable methodologies and tools. Also, model validation is not a discrete step in the simulation process. It needs to be applied continuously from the formulation of the problem to the implementation of the study findings as a completely validated and verified model does not exist. Validation and verification process of a model is never completed. As the model developers are inevitably biased and may be concentrated on positive features of the given model, the third party approach (board of experts) seems to be a better solution in model validation. Further, considering the

ranges that simulation studies cover (from small models to very large-scale simulation models), further research is needed to determine with respect to the size and type of simulation study, i.e. which model validation approach should be used, how should model validation be managed, and what type of support system software for model validation is needed. A final strand of research is validating large-scale simulations that combine different simulation (sub) models and use different types of computer hardware such as in currently being done in Higher Level Architecture. A number of these issues need to deepen research, e.g. how does one verify that the simulation clocks and event (message) times (timestamps) have the same representation (floating point, word size, etc.) and validate that events having time ties are handled properly.

Research Challenge 5.2 - Collaborative model simulations and scenarios generation

Definition. This methodology encompasses participation of all stakeholders in the policy-making process through the implementation of online-based easy-to-use tools for all the levels of skills. Decision-making processes have to be supported with meaningful representations of the present situations along with accurate simulation engines to generate and evaluate future scenarios. Instrumental to all this is the possibility to gather and analyze huge amounts of relevant data and visualize them in a meaningful way also for an audience without technical or scientific expertise. Citizens should also be allowed for probing and real-time data collection for feeding simulation machines at real time, and/or contributing by mean of some sort of online platform. 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. In short scenario simulators and decision support tools. In this framework it is highly important to launch new research directions aimed at developing 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. The weakest form of involvement is feedback to the session facilitator, similar to the conventional way of modelling. Stronger forms are proposals for changes or (partial) model proposals. In this particular approach the modelling process should be supported by a combination of narrative scenarios, modelling rules, and e-Participation tools (all Integrated via an ICT e-Governance platform): so the policy model for a given domain can be created iteratively using cooperation of several stakeholder groups, such as decision makers, analysts, companies, civic society, and the general public.

Relevance and applications in policy making. Clearly the collaboration of several individuals in the simulation and scenario generation allows for policies and impact thereof to be better understood by non-specialists and even by citizens, ensuring a higher acceptance and take up. Furthermore, as citizens have the possibility to intervene in the elaboration of policies, user centricity is achieved. On the other hand, modelling co-creation has also other advantages: no person typically understands all requirements and understanding tends to be distributed across a number of individuals; a group is better capable of pointing out shortcomings than an individual; individuals who participate during analysis and design are more likely to cooperate during implementation.

Technologies, tools and methodologies.

There are several tools and methodologies which are currently used.

- United Nations Global Policy Model (GPM)[69]: this is a tool for investigation of policy scenarios for the world economy. The model is intended to trace historical developments and potential future impacts of trends, shocks, policy initiatives and responses over short, medium and long-term timescales, in the view to provide new insights into problems of policy design and coordination. Recently, the model has been applied to the assessment of possible policy scenarios and implication for the world economy in a post-Brexit setting.
- The European Central Bank New Area-Wide Model (NAWM)[70]: dynamic stochastic general equilibrium model reproducing the dynamic effects of changes in monetary policy interest rates observed in identified Variable Autoregression Models (VARs). The building blocks are: agents (e.g. households and firms), real and nominal frictions (e.g. habit formation, adjustment costs), financial frictions (domestic and external risk premium), rest-of-World block (SVAR). It is estimated on time series for 18 key macro variables employing Bayesian inference methods. The model is regularly used for counterfactual policy analysis**;**
- TELL ME Model (Badham et al. 2018): this a prototype agent-based model, developed within the scope of the European-funded TELL ME project, intended to be used by health communicators to understand the potential effects of different communication plans under various influenza epidemic scenarios. The model is built on two main building blocks: a behaviour model that simulates the way in which people respond to communication and make decisions about whether to vaccinate or adopt other protective behaviour, and an epidemic model that simulates the spread of influenza**;**
- Global epidemic and mobility model (GLEAM)[71]: big data and high performance computing model combining real-world data on populations and human mobility with elaborate stochastic models of disease transmission to model the spread of an influenza-like disease around the globe, in order to be able to test intervention strategies that could minimize the impact of potentially devastating epidemics. An interesting application case quantification of the risk of local Zika virus transmission in the continental US during the 2015-2016 ZIKV epidemic;
- SAFFIER II[72]: this is a macro-economic model developed and used at the Dutch Bureau for Economic Policy Analysis, which delivers economic analysis and forecasts for the government of the Netherlands. SAFFIERII is part of a family of models with a history of several decades.
- 2050 calculator[73]. This is an accounting-type model of carbon use developed by the UK Department for Energy and Climate change for internal and external use. The department is responsible for energy security, affordable energy supplies and climate change mitigation in the UK. The 2050 calculator was developed by a team of about five model developers and is accessible to the general public;
- Simulogue: this tool, developed by Dutt et al. (2019), is designed as a platform for integrated governance through a facilitated dialogue between various stakeholders involved with

[69] https://debt-and-finance.unctad.org/Pages/GPM.aspx
[70] https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp944.pdf

[71] http://www.gleamviz.org/

[72] https://www.cpb.nl/sites/default/files/publicaties/download/doc217.pdf

[73] http://2050-calculator-tool.decc.gov.uk/#/home

governing the city of Chennai (India). The dialogue is based on various future scenarios that each stakeholder develops and is able to negotiate with their peers. A futures-based approach helps to improve decision making by facilitating the integration of diverse public institutions and collaboration between stakeholders and by incorporating intangible data with regards their interaction and decision-making into the decision support system. Further, the tool enables scenario-based planning in the view to explore different situations.

A final example, developed by the team led by Vittorio Loreto, member of our expert committee, is the CityChrone++, which is one of the instantiations of a larger platform dubbed what if-machine (link to whatif.caslparis.com), aimed at providing users with tools to assess the status of our urban and inter-urban spaces and conceive new solutions and new scenarios. The platform integrates flexible data analysis tools with a simple scenario simulation platform in the area of urban accessibility, with a focus on human mobility. Human mobility in cities is driven by several factors, featuring a complex interplay between socio-economic conditions, personal inclinations and needs, the urban environment itself and the status of the transportation systems. Individuals may change their behavior depending on the particular time of the day, on which activities they have to perform, and on the quality of the transportation system itself. On the other hand, individual inclinations to adopt public transportation vs. private vehicles might have a tremendous impact on the future of urban environments. It is thus necessary a better understanding of the behavior of individuals during their daily trips, what drives their choices and which statistical patterns are related to them. This framework allows to parallel the platform with effective modelling schemes, key for the generation and the assessment of new scenarios. Spefically, the platform allows to study the role of current infrastructures and services (knockout, variations..), interplay of different transport modalities, the easy implementation of new scenarios (new infrastructures, new business models, etc.); and a fast impact assessment over different perspectives (economic, social, infrastructural, etc.). An example of scenario development is provided in Figure .

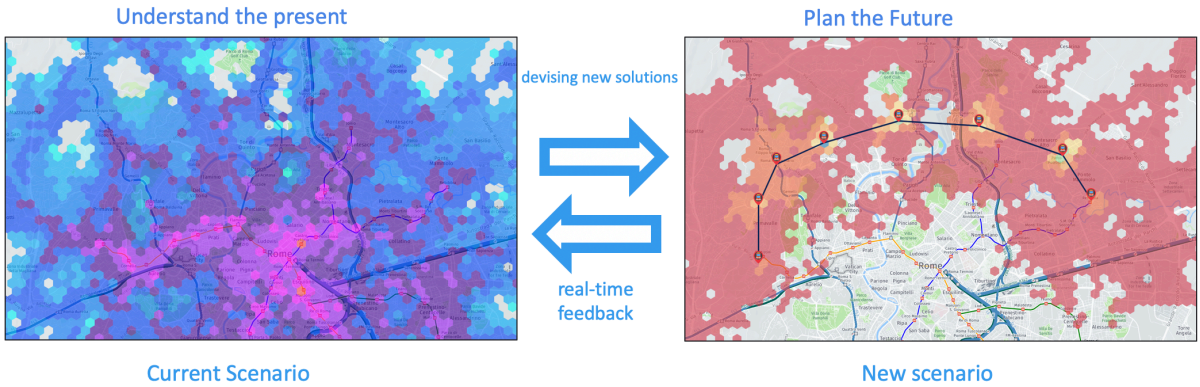

Figure 14 – Scenario Development (source: Vittorio Loreto)

The tool allows also a cross-city comparison, as displayed in.

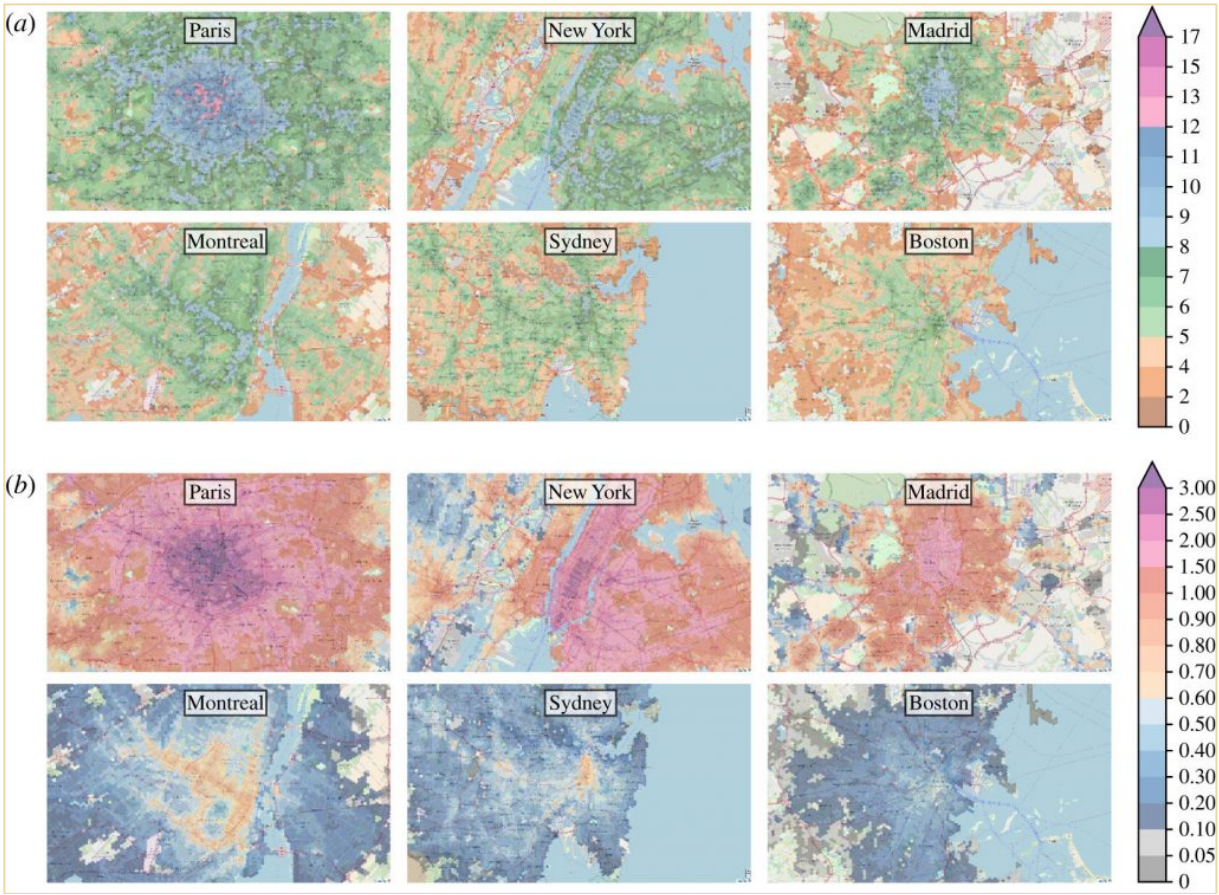

Figure 15 - Maps of the velocity score and the sociality score for six different cities: Paris, New York, Madrid, Montreal, Sidney and Boston (source: Biazzo et al. 2019)

Research perspectives. From the generic point of view of modelling and simulation, an interesting field of research is the combination of simulation models (based on processes) and machine learning models (based on data), in order to increase the potentiality in the analysis. This could be done for instance by applying machine learning prior to the simulation to process input data in order to make it usable for the simulation model. An example would be to develop data-driven decision heuristics that agents can apply. Another approach would be to embed machine learning within the simulation, by training machine learning algorithms on specifics of the simulation. A final approach would be to use machine learning algorithms on the output of the simulations.[74] On a similar note, another interesting domain of research is machine readable engineering and system models. In fact currently many system models are not machine-readable. Engineering models on the other hand are semi-structured because digital tools are increasingly used to engineer a system. Research and innovation in this area of work will assure that machine learning algorithms can leverage system know-how that today is mainly limited to humans. Linked data will facilitate the semantic coupling of know-how at design and implementation time, with discovered knowledge from data at operation time, resulting in self-improving data models and algorithms for machine learning (Curry et al. 2013).

[74] https://www.benjamin-schumann.com/blog/2018/5/7/time-to-marry-simulation-models-and-machine-learning

Apart from this technical advancement, there is the need to develop tools allowing policy makers to have meaningful representations of the present situations along with accurate simulation engines to generate and evaluate future scenarios. 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. In short scenario simulators and decision support tools. In this framework it is highly important to launch new research directions aimed at developing 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. A possible route forward is group model building and systems thinking, focusing on models when tackling a mix of interrelated strategic problems to enhance team learning, foster consensus, and create commitment: although people have different views of the situation and define problems differently, this current field of research shows that this can be very productive if and when people learn from each other in order to build a shared perspective. Some other perspectives include the definition of frameworks allowing even "low-skilled" Citizens to provide their contribution (even If in a discursive way) to the modelling process, the design of more intuitive and accessible Human-Computer Interfaces, and the development of online tools for collaborative model development, such as the aforementioned CityChrone++.

Research Challenge 5.3 - Integration and re-use of modelling schemes

Definition. This research challenge seeks to find the way to model a system by using already existing models or composing more comprehensive models by using smaller building blocks, either by reusing existing objects/models or by generating/building them from the very beginning. Therefore, the most important issue is the definition/identification of proper (or most apt) modelling standards, procedures and methodologies by using existing ones or by defining new ones. Further to that, the present challenge calls for establishing the formal mechanisms by which models might be integrated in order to build bigger models or to simply exchange data and valuable information between the models. Finally, the issue of model interoperability as well as the availability of interoperable modelling environments should be tackled, as well as the need for feedback-rich models that are transparent and easy for the public and decision makers to understand.

Relevance and applications in policy making. In systems analysis, it is common to deal with the complexity of an entire system by considering it to consist of interrelated sub-systems.[75] This leads naturally to consider models as consisting of sub-models. Such a (conceptual) model can be implemented as a computer model that consists of a number of connected component models (or modules). Component-oriented designs actually represent a natural choice for building scalable, robust, large-scale applications, and to maximize the ease of maintenance in a variety of domains. An implementation based on component models has at least two major advantages. First, new models can be constructed by coupling existing component models of known and guaranteed quality with new component models. This has the potential to increase the speed of development. Secondly, the forecasting capabilities of several different component models can be compared, as opposed to compare whole simulation systems as the only option. Further, common and frequently used functionalities, such as numerical integration services, visualization and statistical ex-post analyses tools, can be implemented as generic tools and developed once for all and easily shared by model developers.

[75] See for instance Cilliers 2002

Technologies, tools and methodologies. An interesting case is the Big Data Test Infrastructure under development within the scope of the Connecting Europe Facility of the European Commission. The CEF BDTI building block provides virtual environments that are built based on a mix of mature open source and off-the-shelf tools and technologies. The building block can be used to experiment with big data sources and models and test concepts and develop pilot projects on big data in a virtual environment. Each of these environments are based on a template that supports one or more use cases. These templates can be deployed, launched and managed as separate software environments. Specifically, the Big Data Test Infrastructure will provide a set of data and analytics services, from infrastructure, tools and stakeholder onboarding services, allowing European public organisations to experiment with Big Data technologies and move towards data-driven decision making. Applicability of the BDTI includes descriptive analysis, Social Media Analysis, Time-series Analysis, Predictive analysis, Network Analysis, and Text Analysis. Specifically, BDTI allows public organizations to experiment with big data sources, methods and tools; launch pilot projects on big data and data analytics through a selection of software tools, acquire support and have access to best practice and methodologies on big data; share data sources across policy domains and organisations. The BDTI architecture includes mainly three parts: the software stack (Governance & Security, Data Ingestion, Data Elaboration, and Data Consumption), the infrastructure, and the different data sources to be used by users.

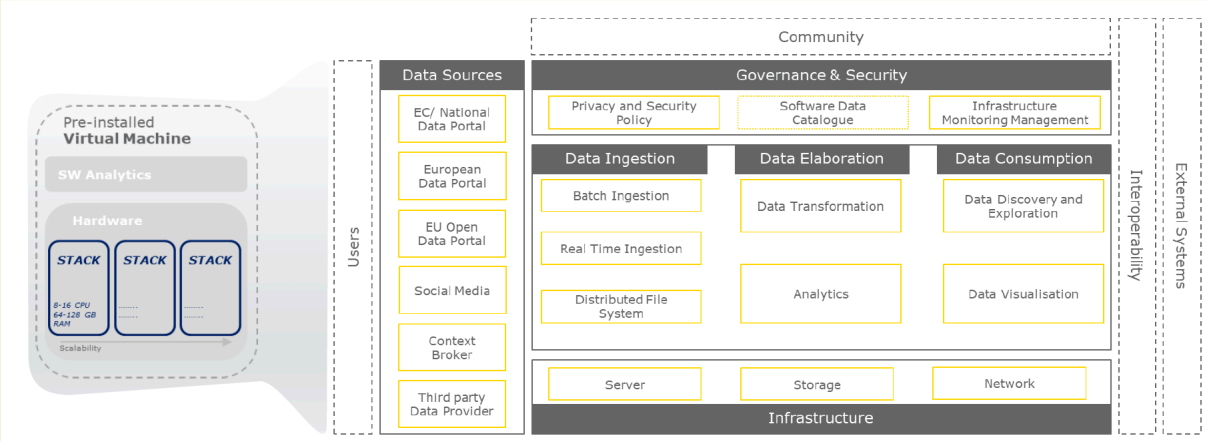

Figure 16 – BDTI Architecture (source: Carbone and de Schouwer 2019)

Users can bring their own data, and make use of the platform as a service.
An interesting application and success story is the one related to the European Big Data Hackathon 2019 carried out by EUROSTAT, with the aim to modernise statistics by mean of automated data collection and more accurate indicators to better support policy decisions. In this respect, it was found that a possible solution would be to experiment with big data from mobile devices to create smart surveys for more accurate statistical indicators.

Research Perspectives**.** Current research, as well as previous research, is still struggling with the problem of different models integration. At present, due to the plethora of different modelling/simulation environments/suites, as well as to differences at the scientific field level, many competing file formats exist. It is possible that vendors perceive the modelling practice as a very small market niche (as the users stem mainly from Academia and to a very small extent from private companies where a Decision Support Systems is used, what is more the Public Administration share is negligible) and therefore are reluctant to introduce interoperable features. Also, current research, as well as previous research, has only recently begun to explore the following issues: open-source modelling and simulation

environments (there are open environments that are rising in importance in the research community, albeit in most cases they only provide the possibility to implement and simulate a model according to the modelling methodology they refer to)[76]; communication of data among models developed in different proprietary (or open) environments by depending on third party solutions; open visualisation of results stemming from model simulation (Zolotas et al. 2019). Regarding future research, one important aspect is the definition of standard procedures for model composition/decomposition (Bae and Moon 2015), e.g. how to deductively pass from a macro-description of models to the definition of its building-blocks, how to inductively conceive a progressive composition of bigger models by aggregating new modules as soon as they are needed or by expanding already existing objects. Another aspect is the proposition of a minimum set of archetypical structures, building blocks or molecules that might be used according to the proper level of decomposition of the model (e.g. systemic archetypes, according to the Systems Thinking/System Dynamics approach, might be useful to describe the overall behaviour thanks to the main variables in the system to be modelled at a macro-to-middle level). The procedures to implement, validate and redistribute any further improvement of these "minimal" objects should be investigated. Then, the definition of open modelling standards, as the basis for interoperability, that is defining common file formats and templates, which would allow the models to be opened, accessed and integrated into every (compliant) model-design and simulation environment. Further, interoperability, also intended in terms of Service Oriented Architectures (Hosseini et al. 2014), as well as the definition and implementation of model repositories (and procedures to add new objects to them), even if they are restricted to hosting models developed according to a specific methodology (Agent Based, System Dynamics, Event Oriented, Stochastic, Hybrid, etc.). Finally, the definition and implementation of new relationships that are created when two models are integrated: all possible important relationships resulting from a model integration/composition should be identified and eventually included in the new deriving integrated model.

5.3.6 Research Challenges on Data Visualization

Research Challenge 6.1 - Automated visualization of dynamic data in real time

Definition. Due to continuing advances in sensor technology and increasing availability of digital infrastructure that allows for acquisition, transfer, and storage of big data sets, large amounts of data become available even in real-time. Since most analysis and visualization methods focus on static data sets, adding a dynamic component to the data source results in major challenges for both the automated and visual analysis methods. Besides typical technical challenges such as unpredictable data volumes, unexpected data features and unforeseen extreme values, a major challenge is the capability of analysis methods to work incrementally. Furthermore, scalability of visualization in face of big data availability is a permanent challenge, since visualization requires additional performances with respect to traditional analytics in order to allow for real time interaction and reduce latency. Finally, visualization is largely a demand-and design-driven research area. In this sense one of the main challenges is to ensure the multidisciplinary collaboration of engineering, statistics, computer science and graphic design.

Based on the Classification of types of big data developed in 2013 by UNECE, there are two main typologies of data that can potentially be collected and visualized in real time: human-sourced information from social network: social networks (Facebook, Twitter, Tumblr etc.), blogs and

[76] See for instance http://sysdyn.simantics.org

comments, personal documents, internet searches, email, mobile data content such as text messages, user generated maps; machine-generated data from Internet of Things: data from sensors, such as fixed sensors (e.g. home automation, weather/pollution sensors, traffic sensors/webcam), mobile sensors (mobile phone location, location of vehicles and planes); satellite data (topographic, thermal, surveillance, meteorological); data from computer systems, like logs and webs logs.

Relevance and applications in policy making. Visualization of dynamic data in real time allows policy makers to react timely with respect to issues they face. An example can be given by movement data (e.g., road, naval, or air-traffic) enabling analysis in several application fields (e.g., landscape planning and design, urban development, and infrastructure planning). In this regard, it helps in identifying problems at an early stage, detect the "unknown unknown" and anticipate crisis: visual analytics of data in real time are for instance largely used in the intelligence community because they help exploiting the human capacity to detect unexpected patterns and connections between data.

Technologies, tools and methodologies. Methodologies for bringing out meaningful patterns include data mining, machine learning, and statistical methods. Tools for management and automated analysis of data streams include: CViz Cluster visualisation, IBM ILOG visualisation, Survey Visualizer, Infoscope, Sentinel Visualizer, Grapheur2.0, InstantAtlas, Miner3D, VisuMap, Drillet, Eaagle, GraphInsight, Gsharp, Tableau, Sisense, SAS Visual Analytics, Zoho Reports. The latter are the most interesting: Tableau allows non-technical users to create interactive, realtime visualizations in minutes[77]; Sisense is a business intelligence tool allowing non-technical users to combine multiple data sets, customize dashboards and generate data visualizations[78]; SAS Visual Analytics is a form of inquiry in which data that provides insight into solving a problem is displayed in an interactive and graphical fashion (Rose 2014); Zoho Reports is an online business application present in cloud[79]. Apart from acquiring and storing the data, great emphasis must be given to the analytics and DSS algorithms that will be used.

Following (Toasa et al. 2018), there is a set of visualization techniques which are suitable for real time data, including autocharting, correlation matrix, network diagram, and Sankey diagrams. Moreover, according to Toasa et al. 2018 an important feature to set up an actual automatic dashboard that reacts when some data is entered into the database in real time, is the use of real time communication technologies such as:

- Redis: An open source (BSD licensed), in-memory data structure store, used as a database, cache, and message broker[80].
- Node.js: this a platform built on Chrome's JavaScript runtime for easily building fast and scalable network applications[81].
- Socket.io: A JavaScript library for real-time web applications that enables real-time, bi-directional communication between web clients and servers[82].

A final interesting application of is provided by Buschmann et al. (2015), who developed a technique for visualizing massive 3D movement trajectories. Specifically, their technique allows to visualizes real-

[77] https://www.tableau.com/solutions/topic/business-dashboards
[78] http://technologyadvice.com/products/sisense-reviews/
[79] http://technologyadvice.com/products/zoho-reviews/
[80] https://redis.io/documentation
[81] https://nodejs.org/es/
[82] https://socket.io/

time simulated movement data by individual attributed trajectories or by aggregated density maps, facilitating spatial reasoning with respect to different application fields such as landscape planning and design, urban development, environmental analysis and simulation, risk and disaster management, as well as logistics and transportation. An example of real-time visualization of massive air-traffic trajectories is provided in Figure .

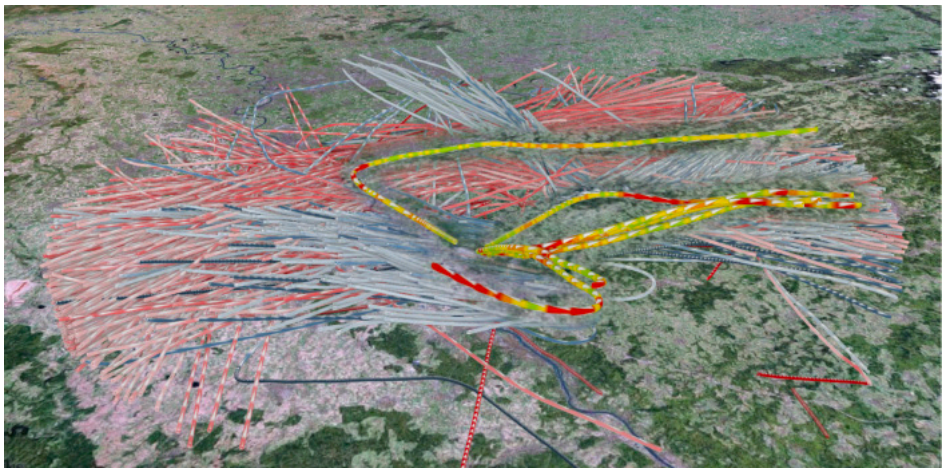

Figure 17 - Visualization of air-traffic trajectories encoding acceleration in color (Source: Buschmann et al. 2015)

Research perspectives. Concerning automated visualization of dynamic data in real time, most of the research carried out is obviously related to the elaboration of visualization interfaces able to display a large amount of data in an understandable way given the human cognitive capability, as well as to the ability to treat and visualize data that are not structured nor processed. More specifically, analysis of such data streams is an important challenge, since the sheer amount of data often does not allow to record all the data at full detail, effective compression and feature extraction methods are needed to manage the data. Furthermore, it is crucial to provide analysis techniques and metaphors that are capable of analyzing large real time data streams in time, and to present the results in a meaningful and intuitive way. The main research perspectives are the following (see inter al. Bikakis 2018; Marks et al. 2014; Du et al. 2017, Sakr and Zomaya 2019; Keim et al. 2009, Wong et al. 2012, Donalek et al. 2014):

- Visualisations techniques that allow interactive exploration techniques such as context and focus, in order for the user to be able to see the object of primary interest presented in full detail while at the same time achieving an overview of all the surrounding information or context available;
- Visualization interfaces for user assistance and personalization, which encourage user comprehension and provide customization capabilities to different user-defined exploration scenarios and preferences according to the analysis needs;
- Development of mobile visualization tools allowing to display data in real time, as well as allowing to display simultaneous multiple visualisation;
- Integration of visualisation with comments stemming from blogs, websites and wikis, and including other meta information such as semantics, data quality, and provenance;
- Develop an evaluation framework for visualisation effectiveness as well as for impact evaluation of visualization on policy choices.
- Detection of bias and signalling in visualisation output
- Development of immersive visualization with virtual reality resulting in a better perception of data scape geometry and more intuitive data understanding

- Development of large-scale visualization based on adaptive semantic frameworks
- Automated collection, generation and visualization of audio and video (animations)
- Development of efficient and scalable techniques supporting the interaction with a large number of objects datasets, while at the same time maintaining a quick system response;
- In situ visual analytics, i.e. visualization of data that are still in memory in order to be able to cope with data amount going beyond the petascale;
- Provision of effective data abstraction mechanisms is necessary for addressing problems related to visual information overloading and visual scalability;
- Provision of data presentation functionalities such as abstract visualization, highly scalable data projection, dimension reduction, high-resolution displays, and power wall displays;
- Visualization techniques copying with high rate of image change;
- Techniques to detect, analyse and extract semantic meta data from heterogeneous sources, and to integrate heterogenous data sources in general;
- Techniques for preprocessing raw data: up- and down-sampling, rounding and weighting, data migration and parsing, aggregation and combining, data cleaning, data reduction and compression, data enrichment, in order to be able to manage large amount of data

Research Challenge 6.2 - Interactive data visualization

Definition. With the advent of Big Data simulations and models grow in size and complexity, and therefore the process of analysing and visualising the resulting large amounts of data becomes an increasingly difficult task. Traditionally, visualisations were performed as post-processing steps after an analysis or simulation had been completed. As simulations increased in size, this task became increasingly difficult, often requiring significant computation, high-performance machines, high capacity storage, and high bandwidth networks. In this regard, there is the need of emerging technologies that addresses this problem by "closing the loop" and providing a mechanism for integrating modelling, simulation, data analysis and visualisation. This integration allows a researcher to interactively perform data analysis while avoiding many of the pitfalls associated with the traditional batch / post processing cycle. This integration also plays a crucial role in making the analysis process more extensive and, at the same time, comprehensible.

Relevance and applications in policy making. Policy makers should be able to independently visualize results of analysis. In this respect, one of the main benefits of interactive data visualization is basically to generate high involvement of citizens in policy-making. One of the main applications of visualization is in making sense of large datasets and identifying key variables and causal relationships in a non-technical way. Similarly, it enables non-technical users to make sense of data and interact with them.[83] Secondly, it helps to understand the impact of policies: interactive visualization is instrumental in making evaluation of policy impact more effective. Interesting applications include:

- Demographics visualisations, allowing stakeholders and decision makers to have a clear picture of the data and of their trends over time. Visualisation of demographic data make easier the design and evaluation of various policies, as there is no need to dig through acres of numbers. In fact, advanced algorithms are able to create figures and illustrations easy to interpret;

[83] See for instance Vornhagen 2018

Legal Arguments visualisation: text analysis, argumentation mappings and visualisation algorithms can be applied to legal documents in order to simplify legislation making it more accessible and comprehensible to the general public;

- Discussion Arguments visualisation, making use of visualisation techniques for visualizing the flow of a discussion that include various arguments, in order to instantly get awareness of the topics discussed, as well as of the arguments and the support such arguments gain. In this view visualisation supports all interested stakeholders to understand the flow of a discussion, which is presented to them in a structured and interactive format, avoiding numerous discussion threads;
- Geovisualisation, which is based on the provision of theory, tools and methods for visual analysis, synthesis, exploration and representation of geographical data and information in order to derive problem specific models and design task specific maps for incorporating geographical knowledge into planning and decision making;
- Advanced visualisation applications used for security and national defense. In this fields, software advances are being led both on the military and on the corporate front. In fact business organizations also have urgent information visualisation requirements that support their business intelligence and situational awareness capability, data mining and reporting requirements. In this view many of the software innovations are being targeted at financial and corporate requirements, but are also applicable to the defense domain due to common data mining and information visualisation challenges.

Technologies, tools and methodologies. Visualisation tools are still largely designed for analyst and are not accessible to non-experts. Intuitive interfaces and devices are needed to interact with data results through clear visualisations and meaningful representations. User acceptability is a challenge in this sense, and clear comparisons with previous systems to assess its adequacy. Furthermore, a good visual analytics system has to combine the advantages of the automatic analysis with interactive techniques to explore data. Behind this desired technical feature there is the deeper aim to integrate the analytic capability of a computer with the abilities of the human analysis. An interesting approach would be to look into two, or even three, tiers of visualisation tools for different types of users: experts and analysts, decision makers (which are usually not technical experts but must understand the results, make informed decisions and communicate their rationale), and the general public (Vornhagen et al. 2019).. Visualisation for the general public will support buy-in for the resulting policies as well as the practice of data-driven policy making in general. Tools available on the market include imMens system, BigVis package for R, Nanocubes, MapD, D3.js, AnyChart, and ScalaR projects, who all use various database techniques to provide fast queries for interactive exploration. According to Wang et al. 2015, the step of interacting visualization are the following: A) interactive selection of data entities or subset or part of whole data or whole data set according to the user interest; B) Linking of relating information among multiple views; C) Filtering the amount of information for display; D) Rearranging the spatial layout of the information. An example of interactive visualization is provided in Figure .

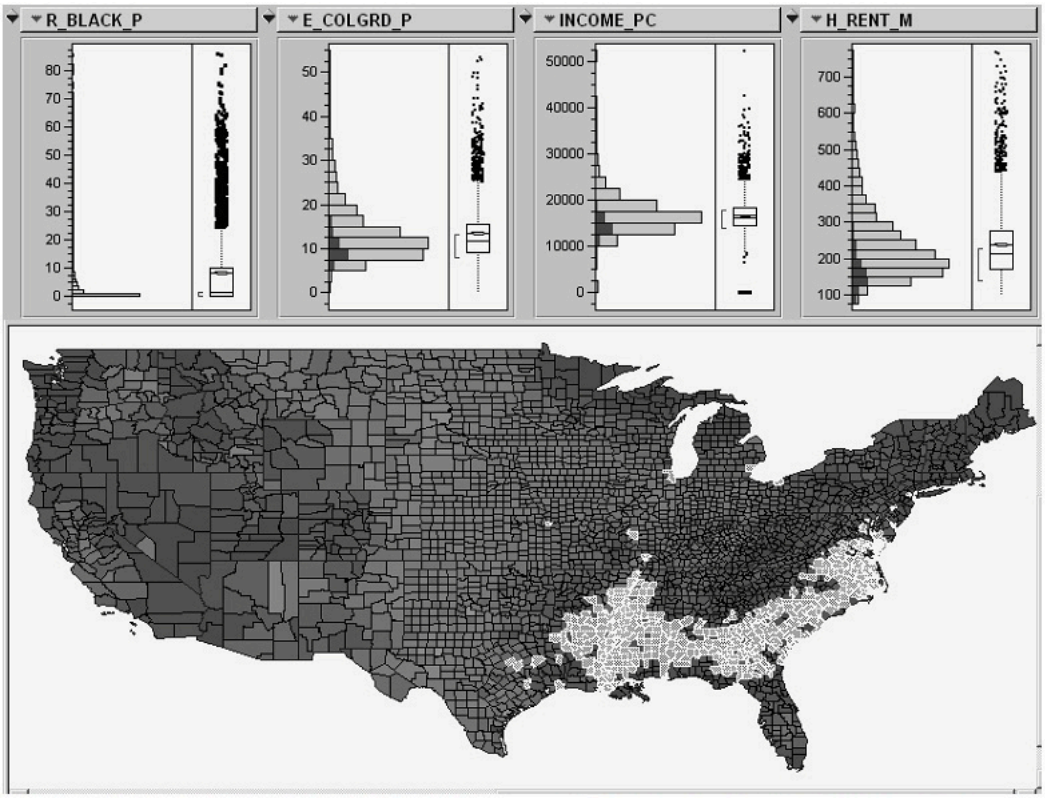

Figure 18 – Example of interactive visualization (Source: Khan & Khan 2011)

A very interesting project with clear applications in the field is PoliVisu, which provides a suite of advanced data processing and visualisation tools coupled with a bespoke methodology for introducing open (geo) data into the policy making lifecycle helps cities overcome barriers to big data use and enables them to leverage the benefits of data analytics to build stakeholder engagement. The PoliVisu tools enable cities to test a variety of policy hypotheses with stakeholders using local data sets (for example road sensor data on traffic flows, historic accident data, traffic light data, pedestrian data etc.) to visually simulate potential impacts. This opportunity to experiment with policy options ensures cities can explore complex systemic urban problems, which require innovative thinking to develop transformative and sustainable solutions, without the need to deploy multiple and costly test pilots. An example of application of the visualization suite allows decision makers in Flanders to discover hotspots of traffic accidents (e.g. accidents in certain hours of the day or on certain weekdays, accidents with certain accident severity, accidents nearby schools etc.) and thus helps to identify most risky areas where to apply specific traffic management or security measures. The mapping is available in Figure .

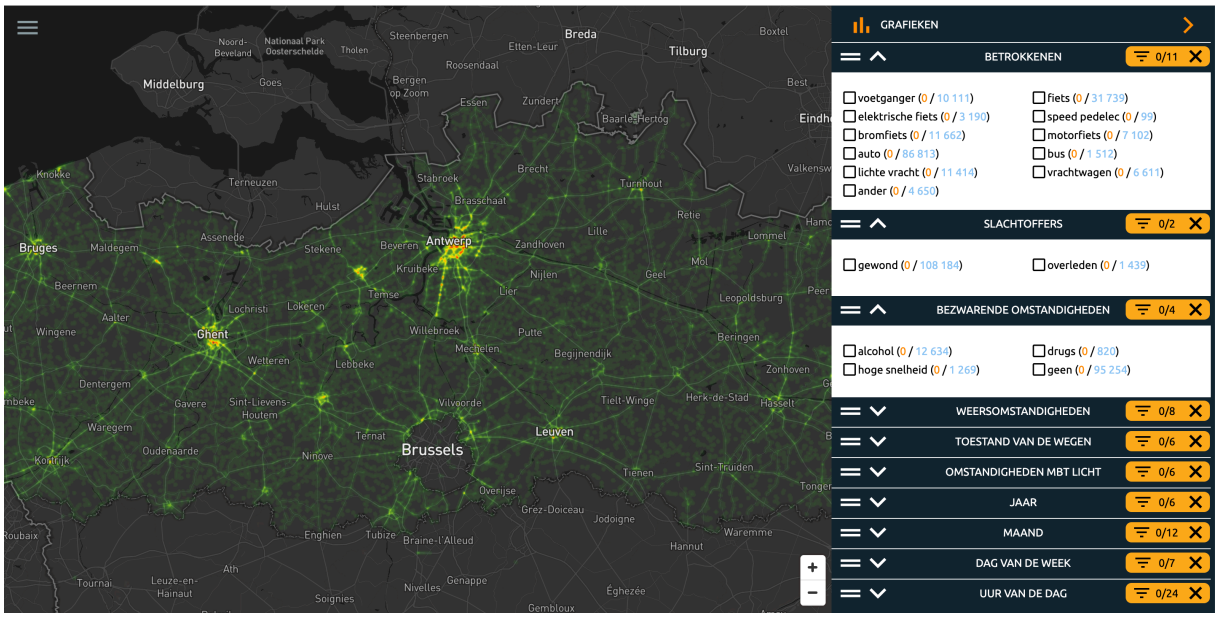

Figure 19 – Hotspots of Traffic Accidents in Flanders

Research perspectives. For what concerns the research perspectives, there are several fields in which research is ongoing and/or will be carried out in the foreseeable future (inter al. Wong et al. 2012; Bikakis 2018; Bikakis et al. 2015; Khan et al 2014; Donalek et al. 2014; Endert et al. 2017). Some general fields regard the integration of visual analytics with opinion mining and participatory sensing, and the research on visualisation as a way to provide (persuasive) feedback and change in attitudes, opinions, behaviours, as well as a medium for grassroots/crowd-sourced participation, collaboration on data-related issues. Further, interesting research is carried out on assessing the efficiency of the visualisation techniques to enable interactive exploration interaction techniques such as focus & context, as well as on the elaboration of evaluation framework for visualisation effectiveness and for impact evaluation of visual analytics on policy choices.

Considering instead the relationship between interactive visualization and algorithms and models:

- Development of learning adaptive algorithm in order to explore the users intent, which is basically the capability to automatically change behaviour based on its execution context in order to obtain optimal performances;
- Development of interaction algorithms that are able to incorporate machine recognition of the actual user intent, as well as appropriate adaptation of main display parameters by which the data is presented;
- Techniques and algorithms for creating effective visualisation tools based on perceptual psychology, cognitive science and graphical principles;
- Application of user modelling techniques to visual analytics and developing collaborative platform display interaction between visualisation and policy models;
- Fostering a tighter integration between automatic computation and interactive visualisation;
- Support of the process of constructing data models for prediction and classification;
- Development of online suites for the integration of modelling and visualization; combination of visual analytics and augmented reality;
- and development of techniques allowing a multilevel hierarchy approach in data scalability.

On the other hand, on the realm of advanced interactive visual interfaces:

- Visual interfaces allowing higher synergy between automation and visualisation;
- Intuitive and affordable visual analytics interface for citizens, as well as multimodal interfaces in hostile working environments;
- Natural language processing for highly variable contexts and visual queries;
- Development of mashable and reusable tools for visual analytics, as well as of mobile visual analytics tools.

Conclusion

This deliverable contains the final version of the Big Policy Canvas Roadmap for Future Research Directions in Data-Driven Policy Making. Starting from a set of gaps and research needs in the use of Big Data in policy making, the deliverable defines six main research clusters related to the use of Big Data in policy making. Four of them are built on the Big Data cycle and value chain, while two are transversal at each phase of the cycle. For each research cluster, a set of research challenges is elaborated. The next step, taking place after the end of the project, will be the elaboration of a joint JRC-BDVA Scientific Report building on the roadmap, to be co-authored by Francesco Mureddu (Lisbon Council), Juliane Schmeling (FOKUS), Gianluca Misuraca (Senior Scientist at the JRC Seville and member of the expert committee), and the Big Data Value Association Smart Cities sub-group. The Scientific Report will be first presented at the High-Level Conference on Data Economy, taking place in Helsinki, on 25-26 November 2019.

References

1 Adler P, Falk C, Friedler SA, et al. (2016) Auditing black-box models by obscuring features. arXiv:1602.07043 [cs, stat].
2 Adrian, M. (2011). Information management goes' Extreme': The biggest challenges for 21st century CIOs. *SAS: Business Analytics and Intelligence Software*.
3 AGCOM(217/17/CONS) Big Data Interim Report.
4 Agrawal R and Srikant R (2000) Privacy-preserving data mining. ACM Sigmod Record. ACM, pp. 439–450.
5 Alfaro, C., J. Cano-Montero, J. Gómez, J. M. Moguerza, and F. Ortega. 2013, September. A multi-stage method for content classification and opinion mining on weblog comments. Annals of Operations Research 1–17.
6 Badham, J, Chattoe-Brown, E, Gilbert, N, Chalabi, Z, Kee, F & Hunter, R 2018, 'Developing Agent-Based Models of Complex Health Behaviour', Health and Place, vol. 54, pp. 170-177.
7 Bae, Jang Won & Moon, Il Chul. (2015). LDEF Formalism for Agent-Based Model Development. IEEE Transactions on Systems, Man, and Cybernetics: Systems.
8 Barabási, A. (2003). *How Everything is Connected to Everything Else and What It Means for Business, Science and Everyday Life*. New York: Plume Books, 294pp. ISBN 0452284392.
9 Barocas S (2014) Data mining and the discourse on discrimination.
10 Bartlett, A., Lewis, J., Reyes-Galindo, L., & Stephens, N. (2018). The locus of legitimate interpretation in Big Data sciences: Lessons for computational social science from -omic biology and high-energy physics. Big Data & Society, 5(1), 205395171876883.
11 Bertot, J.C., & Choi H. (2013). Big data and E-government: issues, policies, and recommendations. In: *Proceedings of the 14th Annual International Conference on Digital Government Research*. dg.o. 2013, 17–20 June, Quebec City, pp. 1–10.
12 Bijlsma, R.M., Bots, P.W.G., Wolters, H.A., & Hoekstra, A.Y. (2011). An empirical analysis of stakeholders' influence on policy development. *Ecol. Soc.* 16(1), 51–66.
13 Bikakis, Nikos & Papastefanatos, George & Skourla, Melina & Sellis, Timos. (2015). A Hierarchical Aggregation Framework for Efficient Multilevel Visual Exploration and Analysis. Semantic Web. 8. 10.3233/SW-160226.
14 Bikakis, Nikos. (2018). Big Data Visualization Tools. Encyclopedia of Big Data Technologies, Springer, 2018.
15 Braun, M. T., & Kuljanin, G. (2015). Big Data and the Challenge of Construct Validity. Industrial and Organizational Psychology, 8(04), 521–527.
16 Bruce, T. and Stiefel, D. (2012), The future of governance and the use of advanced information technologies, Futures, Vol. 44, Issue 9, pp. 812–822.
17 Budäus, D., & Buchholtz K. (1997). Konzeptionelle Grundlagen des Controllings in öffentlichen Verwaltungen. *Die Betriebswirtschaft*, 57(3), 322–337.
18 Buelens, B., Daas, P., Burger, J., Puts, M. and van den Brakel, J. (2014). Selectivity of Big Data.
19 Buelens, B.et al. (2014). Selectivity of Big Data. Discussion paper, 2014/11, Statistics Netherlands, The Hague/Heerlen, The Netherlands.
20 Burrell J (2016) How the machine 'thinks:' Understanding opacity in machine learning algorithms. Big Data & Society 3(1): 1–12.
21 Buschmann, S.; Trapp, M.; Döllner, J. Real-time visualization of massive movement data in digital landscapes. In Proceedings of the 16th Conference on Digital Landscape Architecture (DLA 2015), Dessau, Germany, 4–6 June 2015; pp. 213–220.
22 Calders T and Verwer S (2010) Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery 21(2): 277–292.
23 Charalabidis, Yannis & Alexopoulos, Charalampos & Loukis, E.. (2016). A Taxonomy of Open Government Data Research Areas and Topics. Journal of Organizational Computing and Electronic Commerce. 26. 10.1080/10919392.2015.1124720.

24 Charalabidis, Y., Zuiderwijk, A., Alexopoulos, C., Janssen, M., Lampoltshammer, T.J., Ferro, E. (2018)The World of Open Data: Concepts, Methods, Tools and Experiences. Springer, Cham.
25 Chen A (2017) The human toll of protecting the Internet from the worst of humanity. New Yorker, 28 January.
26 Chen, M., Mao, S., & Liu, Y. (2014). Big data: a survey. *Mobile Networks and Applications*, 19(2), 171–209.
27 Chetty, K., Qigui, L., Gcora, N., Josie, J., Wenwei, L. and Fang, C. (2018), Bridging the digital divide: measuring digital literacy, Economics: The Open-Access, Open-Assessment E-Journal, 122018-23, pp. 1-20.
28 Cilliers, P. (2002). Complexity and postmodernism: Understanding complex systems. Routledge.
29 Christakis N., & Fowler J. (2007). The Spread of Obesity in a Large Social Network Over 32 Years. *New England Journal of Medicine,* 357(4): 370-379.
30 Clarke, R. (2016). Big data, big risks. *Information Systems Journal*, 26(1), 77–90.
31 Colecchia, A. & Schreyer P. (2002). ICT Investment and Economic Growth in the 1990s: Is the United States a Unique Case? A Comparative Study of Nine OECD Countries. *Review of Economic Dynamics*. April, 5:2, pp. 408–42.
32 Conneau A, Schwenk H, Barrault L, et al. (2017) Very deep convolutional networks for text classification. In: Proceedings of the 15th conference of the European chapter of the association for computational linguistics, Valencia, Spain, pp. 1107–1116.
33 Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), 1-4.
34 Cuno, S., Bruhns, L., Tcholtchev, N., L√§mmel, P. and Schieferdecker, I. (2019), "Data Governance and Sovereignty in Urban Data Spaces Based on Standardized ICT Reference Architectures", MDPI, No. Vol 4(1).
35 Cuno, S., Bruhns, L., Tcholtchev, N., Lämmel, P. and Schieferdecker, I. (2019), Data Governance and Sovereignty in Urban Data Spaces Based on Standardized ICT Reference Architectures, MDPI, Vol 4(1).
36 Curry, E. (2016). The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches. In: Cavanillas J., Curry E., Wahlster W. (eds) *New Horizons for a Data-Driven Economy*. Springer, Cham.
37 Danaher, J., Hogan, M. J., Noone, C., Kennedy, R., Behan, A., De Paor, A., … Shankar, K. (2017). Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data & Society*.
38 Davenport, T. H., Barth, P., & Bean, R. (2012). How big data is different. *MIT Sloan Management Review*, 54(1), 43–46.
39 Davis M, Kumiega A and Van Vliet B (2013) Ethics, finance, and automation: A preliminary survey of problems in high frequency trading. Science and Engineering Ethics 19(3): 851–874.
40 de França BBN, Travassos GH (2014) Simulation based studies in software engineering: a matter of validity. In: CIbSE/ESELAW. April. Pucón, Chile.
41 Del Vicario, M., Bessi A., Zollo F., Petroni F., Scala A., Caldarelli G., Stanley, E. & Quattrociocchi, W. (2016). The spreading of misinformation online. Proceedings of the National Academy of Sciences, 113 (3), 554-559.
42 Diakopoulos N (2015) Algorithmic accountability: Journalistic investigation of computational power structures. Digital Journalism 3(3): 398–415.
43 DIN SPEC 91357, Reference Architecture Model Open Urban Platform (OUP), available at: https://www.din.de/en/wdc-beuth:din21:281077528?sourceLanguage&destinationLanguage (accessed 25 August 2019).
44 Discussion paper 201411*, Statistics Netherlands*, The Hague/Heerlen, The Netherlands.
45 Donalek, C.; S. G. Djorgovski, A. Cioc, A. Wang, J. Zhang, E. Lawler, S. Yeh, A. Mahabal, M. Graham, A. Drake, et al. Immersive and collaborative data visualization using virtual reality

platforms. In Big Data (Big Data), 2014 IEEE International Conference on, pp. 609–614. IEEE, 2014.

46 Du, F., Cao, N., Lin, Y., Xu, P., & Tong, H. (2017). iSphere: Focus+Context Sphere Visualization for Interactive Large Graph Exploration. *CHI*.

47 Dunleavy, P. (2016). 'Big data' and policy learning. In G. Stoker, & M. Evans (Eds.), *Evidence-based Policy Making in the Social Sciences: Methods That Matter* (pp. 143-151). Bristol: Policy Press.

48 Dutt, V.; Sil, S.; Krishna, H.; Palavalli, B. (2019). Imagining Futures – A generative scenario-based methodology to improve planning and decision-support systems for policymakers. Conference paper. Data for Policy 2019.

49 Endert, A. & Ribarsky, W. & Turkay, Cagatay & Wong, B.L. & Nabney, Ian & Díaz Blanco, Ignacio & Rossi, Fabrice. (2017). The State of the Art in Integrating Machine Learning into Visual Analytics: Integrating Machine Learning into Visual Analytics. Computer Graphics Forum. 10.1111/cgf.13092.

50 Engin, Z; Koshiyama, A (2019) Digital Ethics and Algorithm Assessment. Conference paper. Data for Policy 2019.

51 Engin, Z., Treleaven, P. Algorithmic Government: Automating Public Services and Supporting Civil Servants in using Data Science Technologies, The Computer Journal, Volume 62, Issue 3, March 2019, Pages 448–460

52 Estevez, E. and Janowski, T. (2013), "Electronic Governance for Sustainable Development—Conceptual framework and state of research", GOVERNMENT INFORMATION QUARTERLY, 2013, 94-S109.

53 European Commission (2018), The Digital Economy and Society Index (DESI), available at: https://ec.europa.eu/digital-single-market/en/human-capital (accessed 26.0.2019).

54 Feng S, Banerjee R and Choi Y (2012) Syntactic stylometry for deception detection. In: Proceedings of the 50th annual meeting of the Association for Computational Linguistics, pp. 171–175.

55 Ferro, E., Loukis, E., Charalabidis, Y., & Osella, M. (2013). Policy making 2.0: From theory to practice. *Government Information Quarterly* 30(4):359–368.

56 Floridi L, Fresco N and Primiero G (2014) On malfunctioning software. Synthese 192(4): 1199–1220.

57 Fule P and Roddick JF (2004) Detecting privacy and ethical sensitivity in data mining results. In: Proceedings of the 27th Australasian conference on computer science – Volume 26, Dunedin, New Zealand, Australian Computer Society, Inc., pp. 159–166.

58 Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. *IDC Iview*, 1142, 1–12.

59 Giest, S. (2017). Big data for policymaking: fad or fasttrack? *Policy Sciences*, 50(3), 367-382.

60 Giglietto, F, L Iannelli, L Rossi, A Valeriani, N Righetti, F Carabini, G Marino, S Usai and E Zurovac (2018), "Mapping Italian news media political coverage in the lead-up of 2018 general election", SSRN.

61 Godfrey, P., Jarek Gryz, Piotr Lasek (2016) Interactive Visualization of Large Data Sets. IEEE Trans. Knowl. Data Eng. 28(8): 2142-2157.

62 Gorrell, G., Greenwood, M. A., Roberts, I., Maynard, D., Bontcheva, K. (2018). Twits, Twats and Twaddle: Trends in Online Abuse towards UK Politicians. 12th Int. AAAI Conf. on Web and Social Media (ICWSM).

63 Grimmelikhuijsen, S., Porumbescu, G., Hong, B., & Im, T. (2013). The Effect of Transparency on Trust in Government: A Cross-National Comparative Experiment. *Public Administration Review*, 73(4), 575-586.

64 Groppo, G., & Heck U. (2009). Strategische Neuausrichtung der öffentlichen Verwaltung. Ein Erfolgsfaktor zur Umsetzung der Verwaltungsmodernisierung. *Verwaltung und Management*, 15(5), 271-277.

65 Guess, A, B Nyhan and J Reifler (2018), "Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign", European Research Council 9.

66 Hajian S, Monreale A, Pedreschi D, et al. (2012) Injecting discrimination and privacy awareness into pattern discovery. In: Data mining workshops (ICDMW), 2012 IEEE 12th international conference on, Brussels, Belgium, IEEE, pp. 360–369.

67 Hanger. S., Pfenninger, S., Dreyfus, M., & Patt, A. (2013). Knowledge and information needs of adaptation policy-makers: a European study. *Regional Environmental Change*, 13(1), 91-101.

68 Harris, S. 2015. The social laboratory. Foreign Policy.

69 Head, B. (2008). Three lenses of evidence-based policy. *Australian Journal of Public Administration*, 67(1), 1–11.

70 Henrich, J., Heine, S. J., & al, et. (2010). The weirdest people in the world. Behavioral and Brain Sciences, 33(2–3).

71 Höchtl, J., Parycek, P., & Schöllhammer, R. (2016). Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce*, 26(1-2), 147-169.

72 Hosseini, M., Ahmadi, M., & Dixon, B. E. (2014). A Service Oriented Architecture Approach to Achieve Interoperability between Immunization Information Systems in Iran. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, *2014*, 1797–1805.

73 Hota, C., Upadhyaya, S., & Al-Karaki, J. N. (2015). Advances in secure knowledge management in the big data era. *Information Systems Frontiers*, 17(5), 983–986.

74 Howard, C. (2005). The Policy Cycle: A Model of Post-Machiavellian Policy Making? *Australian Journal of Public Administration*, 64: 3-13.

75 Howard, P.N. & Kollanyi, B (2016). Bots, #Strongerin, and #Brexit: Computational Propaganda During the UK-EU Referendum. SSRN

76 Indaco, B; Monechi, B; and Loreto, V. (2019) General scores for accessibility and inequality measures in urban areas. Royal Society Open Science 6.8: 190979.

77 International Electrotechnical Commission (2017), Edge intelligence, available at: https://www.iec.ch/whitepaper/pdf/IEC_WP_Edge_Intelligence.pdf (accessed 13 September 2019).

78 Janssen, M., & Helbig N. (2015). Innovating and changing the policy-cycle: policy-makers be prepared! *Gov. Inf. Q*.

79 Janssen, M., & Kuk, G. (2016). Big and open linked data (BOLD) in research, policy, and practice. *Journal of Organizational Computing and Electronic Commerce*, 26(1–2), 3–13.

80 Janssen, M., & Van den Hoven, J. (2015). Big and open linked data (BOLD) in government: a challenge to transparency and privacy? *Government Information Quarterly*, 32(4), 363–368.

81 Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Inf. Syst. Manag*. 29(4), 258–268.

82 Jorgenson, D., & Stiroh K. (2000). Raising the Speed Limit: U.S. Economic Growth in the Information Age. *Brookings Papers on Economic Activity*, (1), 125–211.

83 Kaplan, R., & Norton, D. (1992). The Balanced Scorecard: Measures that Drive Performance. *Harvard Business Review*, 70(1), 71-79.

84 Keim D.A., Mansmann F., Stoffel A., Ziegler H. (2009) Visual Analytics. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA

85 Khan, M.; S.S. Khan (2011) Data and Information Visualization Methods and Interactive Mechanisms: A Survey, International Journal of Computer Applications, 34(1), 2011, pp. 1-14.

86 Khan, Nawsher & Yaqoob, Ibrar & Hashem, Ibrahim & Inayat, Zakira & Kamaleldin, Waleed & Alam, Muhammad & Shiraz, Muhammad & Gani, Abdullah. (2014). Big Data: Survey, Technologies, Opportunities, and Challenges. The Scientific World Journal. 2014. 18. 10.1155/2014/712826.

87 Killmeyer J., Crandall R., Crandall, W. (2014). *Principles of Supply Chain Management*. CRC Press.

88 Kim B, Patel K, Rostamizadeh A, et al. (2015) Scalable and interpretable data representation for high-dimensional, complex data. AAAI. pp. 1763–1769.

89 Kim, B.S., Kang, B.G., Choi, S.H., Kim, T.G., 2017. Data modeling versus simulation modeling in the big data era: case study of a greenhouse control system. Simulation 93 (7), 579–594.

90 Kim, B.S.; Kang, B.G.; Choi, S.H.; Kim, T.G. Data modeling versus simulation modeling in the big data era: Case study of a greenhouse control system. Simulation 2017, 93.

91 Kim, J.K. (2011). Parametric fractional imputation for missing data analysis, *Biometrika*, 98, 119–132.

92 Kim, J.K., Berg, E. & Park, T. (2016). Statistical matching using fractional imputation, Surv. Methodol., 42, 19–40.

93 Kim, Jae Kwang and Wang, Zhonglei, "Sampling techniques for big data analysis in finite population inference" (2018). Statistics Preprints. 136.

94 Kitchin R (2016) Thinking critically about and researching algorithms. Information, Communication & Society 20(1): 14–29.

95 Klievink, B., Romijn, B. J., Cunningham, S., & de Bruijn, H. (2016). Big data in the public sector: Uncertainties and readiness. *Information Systems Frontiers: a journal of research and innovation*, 1-17.

96 Kokkinakos P., Markaki O., Koussouris S., Psarras J. (2016) Digital Transformation: Is Public Sector Following the Enterprise 2.0 Paradigm?. In: Chugunov A., Bolgov R., Kabanov Y., Kampis G., Wimmer M. (eds) Digital Transformation and Global Society. DTGS 2016. Communications in Computer and Information Science, vol 674. Springer, Cham.

97 Kolkman, D.A., Campo, P., Balke-Visser, T. et al. Policy Sci (2016) 49: 489.

98 Kostkova P, Brewer H, de Lusignan S, Fottrell E, Goldacre B, Hart G, Koczan P, Knight P, Marsolier C, McKendry RA, Ross E, Sasse A, Sullivan R, Chaytor S, Stevenson O, Velho R and Tooke J (2016) Who Owns the Data? Open Data for Healthcare. Front. Public Health 4:7.

99 Kuhlmann S., & Meyer-Krahmer F. (1995). Practice of Technology Policy Evaluation in Germany: Introduction and Overview. In: Becher G., Kuhlmann S. (eds) Evaluation of Technology Policy Programmes in Germany. Economics of Science, Technology and Innovation, vol 4. Springer, Dordrecht.

100 Lagoze, C. (2014). Big Data, data integrity, and the fracturing of the control zone. Big Data & Society, 1(2), 205395171455828.

101 Le Q and Mikolov T (2014) Distributed representations of sentences and documents. In: Proceedings of the 31st international conference on machine learning, Beijing, China, pp. II-1188–II-1196.

102 Lewis, M. and Pettersson, G. (2009), Governance in Education: Raising Performance.

103 Linders, D. (2012). From E-government to we-government: defining a typology for citizen coproduction in the age of social media. *Gov. Inf. Q*. 29(4), 446–454.

104 Longo, J., Kuras, E., Smith, H., Hondula, D. M., & Johnston, E. (2017). Technology use, exposure to natural hazards, and being digitally invisible: implications for policy analytics. *Policy & Internet*, 9(1), 76-108.

105 Lou Y, Caruana R, Gehrke J, et al. (2013) Accurate intelligible models with pairwise interactions. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. Chicago, USA, ACM, pp. 623–631.

106 Maciejewski, M. (2017). To do more, better, faster and more cheaply: using big data in public administration. *International Review of Administrative Sciences*, *83*(1_suppl), 120–135.

107 Madelin, R. (2015), "SCIENCE AS THE FUEL OF THE PUBLIC POLICY MACHINE", in Wilsdon, J. and Doubleday, R. (Eds.), Future Research Directions for Scientific Advice in Europe, Centre for Science and Policy, Campbridge, pp. 26-32.

108 Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and H.-B. Angela. 2011. Big data: The next frontier for innovation, competition, and productivity. Washington, DC: McKinsey Global Institute.
109 Margetts, H., & Sutcliffe, D. (2013). Addressing the policy challenges and opportunities of Big data. *Policy & Internet*, 5(2), 139–146.
110 Marks, S., Estevez, J.E. & Connor, A.M. (2014) Towards the Holodeck: Fully Immersive Virtual Reality Visualisation of Scientific and Engineering Data. Proceedings of the 29th International Conference on Image and Vision Computing New Zealand.
111 Martin, Kirsten. (2018). Ethical Implications and Accountability of Algorithms. Journal of Business Ethics. 10.1007/s10551-018-3921-3.
112 Mayer-Schönberger, V., & Cukier, K. (2013). *Big data. A revolution that will transform how we live, work and think*. London: John Murray Publishers.
113 McCarthy, N; Fourniol, F. (2019) The Role of Technology in Governance: the Example of Privacy Enhancing Technologies. Conference paper. Data for Policy 2019.
114 McCombs, M. E., and D. L. Shaw. 1972. The agenda-setting function of mass media. Public Opinion Quarterly 36 (2):176–87.
115 Mcintosh, B. S., Alexandrov, G., Matthews, K., Mysiak, J., & van Ittersum, M. (2011). Preface: Thematic issue on the assessment and evaluation of environmental models and software. Environmental Modelling and Software, 26(3), 245–246.
116 Mcintosh, B. S., Giupponi, C., Voinov, A. A., Smith, C., Matthews, K. B., Monticino, M., et al. (2008). Bridging the gaps between design and use: Developing tools to support environmental management and policy. In A. J. Jakeman, A. A. Voinov, A. E. Rizzoli, & S. H. Chen (Eds.), Environmental Modelling, Software and Decision Support: State of the art and new perspective. Amsterdam: Elsevier.
117 Medvedeva M, Kroon M and Plank B (2017) When sparse traditional models outperform dense neural networks: The curious case of discriminating between similar languages. In: Proceedings of the 4th workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), Valencia, Spain, pp. 156–163.
118 Micheli, M., Blakemore, M., Ponti, M., Scholten, H. and Craglia, M. (2018), The Governance of Data in a Digitally Transformed European Society, Second Workshop of the DigiTranScope Project, JRC114711, European Commission.
119 Millard, J. (2015). Open governance systems: doing more with more. *Gov. Inf. Q.*
120 Mittelstadt BD and Floridi L (2016) The ethics of big data: Current and foreseeable issues in biomedical contexts. Science and Engineering Ethics 22(2): 303–341.
121 Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*.
122 Monahan, Torin; Fisher, Jill A. (June 1, 2010). "Benefits of 'Observer Effects': Lessons from the Field". U.S. National Library of Medicine, National Institutes of Health.
123 Moorthy, J., Lahiri, R., Biswas, N., Sanyal, D., Ranjan, J., Nanath, K., & Ghosh, P. (2015). Big Data: Prospects and Challenges. *Vikalpa*, *40*(1), 74–96.
124 Neyland D (2016) Bearing accountable witness to the ethical algorithmic system. Science, Technology & Human Values 41(1): 50–76.
125 OECD (2013). Exploring data-driven innovation as a new source of growth: Mapping the policy issues raised by "big-data", in: OECD: Supporting Investment in Knowledge Capital, Growth and Innovation. *OECD Publishing*, Paris, 319-356.
126 OECD (2014), Impact assessment in STI policies, OECD Science, Technology and Industry Outlook 2014, OECD Publishing, Paris.
127 OECD (2017a) OECD Digital Government Toolkit. 12 Principles.
128 OECD (2017b). Knowledge management in the public and private sectors: similarities and differences in the challenges created by the knowledge-intensive economy.
129 OECD (2018), New approaches in policy design and experimentation, OECD Science, Technology and Innovation Outlook 2018: Adapting to Technological and Societal Disruption,

OECD Publishing, Paris, available at: https://www.oecd-ilibrary.org/sites/sti_in_outlook-2018-16-en/index.html?itemId=/content/component/sti_in_outlook-2018-16-en (accessed 16 September 2019).

130 Olabe, P. B. (2017). Responding to citizen´s need: Public services and trust, in: OECD, Trust and Public Policy. How Better Governance Can Help Rebuild Public Trust, *OECD Publishing*, Paris, 47-65.

131 Oliner, S., & Sichel, D. (2000). The Resurgence of Growth in the Late 1990s: Is Information Technology the Story? *Mimeo, Federal Reserve Board*, February.

132 OpenTracker (2013). Definitions of big data. *OpenTracker*.

133 Ormerod, P. (2010). N squared: public policy and the power of networks. *RSA Pamphlets*, p36.

134 Orseau L and Armstrong S (2016) Safely interruptible agents. Available at: http://intelligence.org/files/Interruptibility.pdf.

135 Osborne, S. P., Radnor, Z., & Nasi, G. (2013). A New Theory for Public Service Management? Toward a (Public) Service-Dominant Approach. The American Review of Public Administration, 43(2), 135–158.

136 Panel for the Future of Science and Technology, Automated tackling of disinformation, European Science-Media Hub, European Parliamentary Research Service, PE 624.278, Brussels, March 2019.

137 Park, S., Kim, J.K. & Stukel, D. (2017). A measurement error model for survey data integration: combining information from two surveys, *Metron,* 75, 345–357.

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

139 Pereira, Gabriela & Parycek, Peter & Falco, Enzo & Kleinhans, Reinout. (2018). Smart governance in the context of smart cities: A literature review. Information Polity. 23. 1-20.

140 Pérez-Rosas V and Mihalcea R (2015) Experiments in open domain deception detection. In: Proceedings of the conference on empirical methods in natural language processing, pp. 1120–1125.

141 Pollanen, R. M. (2016). Linking Strategic Planning and Performance. Measurement in Canadian Public organizations: Does it improve Performance? *Financial Management Institute of Canada*.

142 Quattrociocchi, W., A. Scala, and C. R. Sunstein. (2016). Echo Chambers on Facebook. Social Science Research Network.

143 References

144 Reyna, A.; Martín, C.; Chen, J.; Soler, E.; Díaz, M. On blockchain and its integration with IoT Challenges and opportunities. *Future Gener. Comput. Syst.* **2018**, *88*, 173–190.

145 Rinnerbauer, B., Thurnay, L., Lampoltshammer, T. J. (2018). Limitations of Legal Warranty in Trade of Data. Virkar, S., Parycek, P. Edelmann, N., Glassey, O., Janssen, M., Scholl, H. J., Tambouris, E., Proceedings of the International Conference EGOV-CeDEM-ePart 2018, 3-5 September 2018. Danube University Krems, Austria: 143-151, Edition Donau-Universität Krems.

146 Robinson, D., & Yu, H. (2012). The New Ambiguity of Open Government. UCLA Law Review Discourse, 1-17.

147 Romei A and Ruggieri S (2014) A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review 29(5): 582–638.

148 Rose, S. "Return on Information : The New ROI Getting value from data.," SAS Inst. Inc. U.S.A, 2014.

149 Ruths, D., & Jurgen, P. (2014). Social media for large studies of behavior. *Science*, 1063-1064.

150 Sakr, S.; Zomaya, A. Y. (2019) Encyclopedia of Big Data Technologies. Springer 2019, ISBN 978-3-319-63962-8.

151 Salganik, M. J. (2019). Bit by bit: Social research in the digital age. Princeton: Princeton University Press.

152 Sandvig C, Hamilton K, Karahalios K, et al. (2014) Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and Discrimination: Converting Critical Concerns into Productive Inquiry.
153 Schelling, T. (1969). Models of Segregation. *American Economic Review,* 59 (May), pp. 488-493.
154 Scheufele, D. A. 1999. Framing as a theory of media effects. Journal of Communication 49:103–122.
155 Schintler, L. A., & Kulkarni, R. (2014). Big data for policy analysis: The good, the bad, and the ugly. *Review of Policy Research*, 31(4), 343-348.
156 Schmeling, J., Marx, A. and Kurrek, H., (2019) Evidenzbasiert steuern. Die integrierte Nutzung von Verwaltungsdaten.
157 Selva Rathna and T. Karthikeyan, "Survey on Recent Algorithms for Privacy Preserving Data mining", IJCSIT, Vol. 6 (2) , 2015, 1835-1840, ISSN: 0975-9646.
158 Shulner Tal, Avital; Hartman, Alan; Batsuren, Khuyagbaatar; Kleanthous Loizou, Styliani; Bogina, Veronika; Tsvi, Kuflik; Giunchiglia, Fausto; Otterbacher, Jahna, "“End to End” Towards a Framework for Reducing Biases and Promoting Transparency of Algorithmic Systems" in Proceedings of th 14th International workshop on Semantic and Social Media Adaptatoin and Personalization, Larnaca, Cyprus: UMAP, 2019. Proceedings of: 14th International workshop on Semantic and Social Media Adaptatoin and Personalization, Larnaca, Cyprus, 9-10 June 2019.
159 Simon J (2015) Distributed epistemic responsibility in a hyperconnected era. In: Floridi L (ed.) The Onlife Manifesto. Springer International Publishing, pp. 145–159.
160 Simon, P. (2013). *Too big to ignore: The business case for big data*. Hoboken: Wiley.
161 Snášel, V., Nowaková, J., Xhafa, F., & Barolli, L. (2017). Geometrical and topological approaches to Big Data. Future Generation Computer Systems, 67, 286–296.
162 Social Media Research Group (2016). *Using social media for social research: An introduction*. UK: Government social research profession.
163 Sørensen, E. and Torfing, J. (2007), Theories of Democratic Network Governance, palgrave macmillan.
164 Taleb, Ikbal & Serhani, Mohamed & Dssouli, Rachida. (2018). Big Data Quality Assessment Model for Unstructured Data.
165 Taleb, N. (2007) *The Black Swan: The Impact of the Highly Improbable.* Random House, ISBN 978-1400063512.
166 Taylor L, Floridi L and van der Sloot B (eds) (2017) Group Privacy: New Challenges of Data Technologies, 1st ed. New York, NY: Springer.
167 The Centre for Public Impact (2017). Destination unknown: Exploring the impact of Artificial Intelligence on Government.
168 Thomas, J. C. (2013), Citizen, Customer, Partner: Rethinking the Place of the Public in Public. *Public Administration Review*, 73(6), 786-796.
169 Toasa, R.; M. Maximiano, C. Reis and D. Guevara, "Data visualization techniques for real-time information — A custom and dynamic dashboard for analyzing surveys' results," 2018 13th Iberian Conference on Information Systems and Technologies (CISTI), Caceres, 2018, pp. 1-7.
170 Tufekci Z (2015) Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. Journal on Telecommunications and High Technology Law 13: 203.
171 Tufekci, Zeynep. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In ICWSM ’14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, 2014.
172 Tutt A (2016) An FDA for algorithms. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network.
173 Ubaldi, B. (2013), "Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives", *OECD Working Papers on Public Governance*, No. 22, OECD Publishing, Paris,

174 UNECE (2014). A Suggested Framework for the Quality of Big Data. UNECE, December 2014.
175 United Nations (2007), Public Governance Indicators: A Literature Review.
176 Vaidhyanathan, S. (2018). Antisocial media: How Facebook disconnects us and undermines democracy. Oxford University Press.
177 van Delden, H., Seppelt, R., White, R., & Jakeman, A. J. (2011). A methodology for the design and development of integrated models for policy support. Environmental Modelling and Software, 26(3), 266–279.
178 van Fleur Veenstra, A. and Kotterink, B. (2017), "Data-Driven Policy Making: The Policy Lab Approach", in Parycek, P., Charalabidis, Y., Chugunov, A.V., Panagiotopoulos, P., Pardo, T.A., S√¶b√∏, √ò., Tambouris, E., van Veenstra, A.F. and Kotterink, B. (Eds.), Data-Driven Policy Making: The Policy Lab Approach: Electronic Participation, Cham, Springer International Publishing, pp. 100-111.
179 Vellido A, MartıÅLn-Guerrero JD and Lisboa PJ (2012) Making machine learning models interpretable. In: ESANN 2012 proceedings, Bruges, Belgium, pp. 163–172.
180 Veltri, G. A. (2019). Digital Social Research. Oxford: Polity Press, 2019.
181 Virkar, S., Viale Pereira, G. (2018) Exploring Open Data State-of-the-Art: A Review of the Social, Economic and Political Impacts. Parycek et. al. Electronic Government EGOV2018. Lecture Notes in Computer Science vol.11020. Springer, Cham.
182 Virkar S., Viale Pereira G., Vignoli M. (2019) Investigating the Social, Political, Economic and Cultural Implications of Data Trading. Lindgren I. et al. (eds) Electronic Government. EGOV 2019. Lecture Notes in Computer Science, vol 11685.Springer, Cham.
183 Vornhagen, H. (2018). Effective Visualisation to Enable Sensemaking of Complex Systems. The Case of Governance Dashboard. In Virkar, Shefali, Peter Parycek, Noella Edelmann, Olivier Glassey, Marijn Janssen, Hans Jochen Scholl, and Efthimios Tambouris, eds. Proceedings of the International Conference EGOV-CeDEM-ePart 2018: 3-5 September 2018 Danube University Krems, Austria (pp. 313-321). Edition Donau-Universität Krems, 2018.
184 Vornhagen, H., Davis, B., & Zarrouk, M. (2018). Sensemaking of complex sociotechnical systems: the case of governance dashboards. In Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age. ACM Press.
185 Vornhagen, H., Young, K. and Zarrouk, M. (2019) Understanding My City through Dashboards. How Hard Can It Be?. In Virkar, S., Glassey, O., Janssen, M., Parycek, P., Polini, A, Re, B., Reichstädter, P., Scholl, H.J., Tambouris, E. (eds.) EGOV-CeDEM-ePart 2019: Proceedings of Ongoing Research, Practitioners, Posters, Workshops, and Projects of the International Conference EGOV-CeDEM-ePart 2019. 2-4 September 2019, San Benedetto del Tronto, Italy. (pp. 21-30).
186 Wang, Lidong, Guanghui Wang, and Cheryl Ann Alexander. "Big Data and Visualization: Methods, Challenges and Technology Progress." Digital Technologies 1.1 (2015): 33-38.
187 Wang, X., & Van Wart, M. (2007), When Public Participation in Administration Leads to Trust: An Empirical Assessment of Managers´ Perceptions. *Public Administration Review*, 67(2), 265-278.
188 Weber, M. (1980). *Wirtschaft und Gesellschaft: Grundriss der verstehenden So*ziologie. Mohr, Tübingen.
189 Wigg, K. M. (2002). Knowledge management in public administration. *Journal of Knowledge Management*, 6(3), 224-239.
190 Wilsdon, J. and Doubleday, R. (Eds.) (2015), Future Research Directions for Scientific Advice in Europe, Centre for Science and Policy, Campbridge.
191 Wilsdon, J., Doubleday, R. and Hynard, J. (2015), "FUTURE DIRECTIONS FOR SCIENTIFIC ADVICE IN EUROPE", in Wilsdon, J. and Doubleday, R. (Eds.), Future Research Directions for Scientific Advice in Europe, Centre for Science and Policy, Campbridge, pp. 8-24.
192 Wong, P. C., Shen, H. W., Johnson, C. R., Chen, C., & Ross, R. B. (2012). The top 10 challenges in extreme-scale visual analytics. *IEEE computer graphics and applications*, *32*(4), 63–67.

193 World Bank (2016), World Development Report 2016: Digital Dividends, World Bank, Washington, DC.
194 World Economic Forum (2016), Digital Transformation of Industries, available at: http://reports.weforum.org/digital-transformation/wp-content/blogs.dir/94/mp/files/pages/files/dti-societal-implications-white-paper.pdf (accessed 19 September 2019).
195 Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. Computers in Human Behavior, 98, 166–173.
196 Yang Z, Yang D, Dyer C, et al. (2016) Hierarchical attention networks for document classification. In: Proceedings of the 2016 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480–1489.
197 Yeung, Karen, 'Hypernudge': Big Data as a Mode of Regulation by Design' (May 2, 2016). Information, Communication & Society (2016) 1,19; TLI Think! Paper 28/2016.
198 Zarsky T (2016) The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. Science, Technology and Human Values 41(1): 118–132.
199 Zhang X, Zhao J and LeCun Y (2015) Character-level convolutional networks for text classification. In: Proceedings of the 28th international conference on neural information processing systems, Montréal, Canada, pp. 649–657.
200 Zolotas, A., H. H. Rodriguez, D. S. Kolovos, R. F. Paige and S. Hutchesson, "Bridging Proprietary Modelling and Open-Source Model Management Tools: The Case of PTC Integrity Modeller and Epsilon," 2017 ACM/IEEE 20th International Conference on Model Driven Engineering Languages and Systems (MODELS), Austin, TX, 2017, pp. 237-247.
201 Zuboff, S. (2015), Big Other: Surveillance Capitalism and the Prospects of an Information Civilization, Journal of Information Technology No. 30, pp. 75-89.

6 Annex I - Assets assessment against Needs functionalities (Step 3)

Table 19 - Asset assessment for N-S-1

Functionalities / Assets	F1: Management control system to monitor political targets based on multiple indicators	F2: Definition of clear goals in policy building with balanced targets
African Highland Farmer – the Game	N/A	Gamification may help policy makers to understand better the impact of policies during policy making process and set appropriate clear and balanced targets
Aragon Open Data	Provides a structured access to data form the government that can feed control systems to monitor political targets	N/A
ENAP	Supports impact assessment to verify that the effects of a project correspond to sustainable development in accordance with the German legislation	N/A
GENIX	N/A	N/A
ISO	N/A	N/A
LEED	N/A	Provides green building rating system supporting policies in the scope of Energy and Environmental Design
Smart Start	Innovative techniques to analyse big data from a wide range of sources to achieve beneficial childhood experiences that allow children to grow up safely and child-friendly.	N/A

Table 20 - Asset assessment for N-S-2

Functionalities / Assets	F1: Participative democracy	F2: Improvement of efficiency and effectiveness by transferring to PAs the experiences, wishes and needs of the citizens into administrational needs in the policy making process
Crowdsourcing Through Social Media-The Icelandic Constitution Case	Citizens can contribute directly to the drafting on the constitution	N/A
D-CENT	Provides tools for enabling democratic and participatory processes	N/A
EtherSport: Blockchain Sports Prediction Platform	N/A	N/A
EVOKE	N/A	Citizens can contribute with creative solutions to real life problems through a game
Fix My Street	N/A	Citizens reporting issues in the streets, so the city council can solve it

Functionalities / Assets	F1: Participative democracy	F2: Improvement of efficiency and effectiveness by transferring to PAs the experiences, wishes and needs of the citizens into administrational needs in the policy making process
Ideas for Bristol	N/A	Crowdsourcing site to involve citizens to provide ideas to reshape the city of Bristol
Improve the Neighborhood	N/A	Participation of citizens to report problems or provide ideas to improve the neighbourhood
Inflation Island	N/A	Educational game to check different inflation and deflation scenarios
LiquidFeedback	Tool to propose and vote ideas (digital assembly)	Tool to propose and vote ideas (civic participation)
Lisbon City Hall - Participatory Budgeting	Participatory platform to elaborate budget based on proposals	N/A
Madrid Participa	Participatory budgeting and public input and feedback on a variety of policy and issue areas	N/A
Maryland Budget Game	Game to make proposals on state budget adjustment	N/A
Regulations.gov	Citizens can provide comments on proposed regulations by the US federal administration	N/A
Smart Start	N/A	Collects inputs to improve childhood experiences in different social environments
Thousand Visions	Game to engage stakeholders to define transportation budget for the transportation of the future	N/A
UrbanSim	N/A	Simulation for supporting planning and analysis of urban development, transportation and land use.
Vancouver User Voice	N/A	Ideation process to collect ideas, votes and comments to make the city more environmentally responsible

Table 21 - Asset assessment for N-S-4

Functionalities / Assets	F1: Citizens' cooperation	F2: Transparency
Agora Voting	Secure and transparent digital voting based on blockchain technology	N/A
Aragon Open Data	N/A	Cross domain Open data from Aragon region (Spain)
BDVA labelled I-Spaces	N/A	i-Spaces are Trusted Data Incubators targeted to accelerate take up of data driven innovation in commercial sectors

Functionalities / Assets	F1: Citizens' cooperation	F2: Transparency
Crowdsourcing Through Social Media-The Icelandic Constitution Case	Citizens can contribute directly to the drafting on the constitution	N/A
D-CENT	N/A	Enables citizens to be informed and get real-time notifications about issues that matter to them
energie atlas	N/A	Information to the citizens and companies of the State of Bavaria in Germany in the domain of energy
Fix My Street	Citizens reporting issues in the streets, so the city council can solve it	N/A
Fraunhofer E-Health	N/A	N/A
GovTrack	N/A	Makes information about the United States Congress accessible, understandable, and actionable for public use
Ideas for Bristol	Involves the city's residents in the redevelopment of the city centre	N/A
Improve the Neighbourhood	Involves citizens in the improvement of their cities	N/A
Inflation Island	To participate in policy making, citizens should understand the concepts of the economy. This game shows how inflation affects the economy	N/A
It's Your Parliament	N/A	Unique overview of the votes cast in the European Parliament, where one can easily find and compare voting records of members of the European Parliament (MEPs) and political groups. It is also possible to make own comments and cast own votes
LiquidFeedback	Platform for proposition development and voting	N/A
Lisbon City Hall - Participatory Budgeting	Involves citizens in the improvement of their city, that can take part in budgeting process	N/A
Madrid Participa	Citizen forums and investments agreed between the City Council and the citizens	N/A
Maryland Budget Game	Game that allows users to develop their own proposals for balancing the state budget	N/A
OpenGov.gr	Open calls for the recruitment of public administration officials; Allows electronic deliberation exploring new ways to tackle modern public administration problems	N/A

Functionalities / Assets	F1: Citizens' cooperation	F2: Transparency
Regulations.gov	Makes it easy to submit a comment on proposed regulations and related documents	N/A
SeeClickFix	Enables users to report non-emergency issues in their communities	Community and local government responses are reported and tracked by users
Thousand Visions	To participate in policy making, citizens should understand the concepts of the economy.The game allows the player to determine the taxes, the projects and the priorities.	N/A
Vancouver User Voice	Ideation process to collect ideas, votes and comments to make the city more environmentally responsible	N/A
€CONOMIA - The Monetary Policy Game	To participate in policy making, citizens should understand the concepts of the economy. This game shows for example how key interest rate affects inflation	N/A

Table 22 - Asset assessment for N-S-9

Functionalities / Assets	F1: Information sharing
Aragon Open Data	Open data from different policy domains and departments ready to be used
BehavePlus	Information sharing in US.Forest Service, that leads to a better understanding of fire behaviour
Blockcerts: An open Standard for Blockchain educational certificates	The citizen can share personal data (blockchain-based certificates about civic records, academic credentials, professional licenses, workforce development, etc.)
DCAT Application Profile for Data Portals in Europe (DCAT-AP)	Provides a common specification for describing public sector datasets in Europe to enable the exchange of descriptions of datasets among data portals
Enquete-Kommission "Internet und digitale Gesellschaft"	N/A
Europeana	Website that shares cultural heritage for enjoyment, education and research
Google Fusion Tables	Experimental data visualization web application to gather, visualize, and share data tables
IBM Watson	Makes sense of data to make better decisions
MAPR	Data platform that harnesses, manages, protects data, and powers the next generation of AI and analytics applications that are essential for data-driven transformation
POPVOX	Platform to exchange opinions on political initiatives. Dialogue between US Congress and trade and union organisations, as well as the general public on specific pieces of legislation

Functionalities / Assets	F1: Information sharing
SAHARA Smart analysis	A medical smart analysis platform for health care
Smart Start	Programme in The Netherlands that develops a data-driven and fact-based approach analysing big data from a wide range of sources to estimate the risk to the child's future well-being
UrbanSim	Simulation platform for supporting planning and analysis of urban development that makes use of shared open data of land use, transportation, the economy, and the environment
X-Road	A platform that allows the secure exchange of data in order to provide efficient public services. The tool can write to multiple databases, transmit large data sets and perform searches across several databases simultaneously. It gives a seamless service provision for citizens, given that once the data is updated, all other service providers will automatically also operate with up to date information

Table 23 - Asset assessment for N-O-7

Functionalities / Assets	F1: Legal basis & specifications
Aragon Open Data	Ontology to organise all the information published by the Aragon Government
Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)	Standardisation of data and process decision making to assess measures to rehabilitate prisoners and parolees
ISO	Non-governmental international organisation with membership from 164 national standards bodies
LEED	Provides green building rating system supporting policies in the scope of Energy and Environmental Design
OpenText	N/A
Polish E-Consultations	N/A
Smart City Reference Architecture German Institute for Standardization	DIN standard. Reference architecture Model Open Urban Platform (UOP) for Smart Cities
Solver BI360	N/A
The public safety assessment	Provides a neutral tool, evidence-based, to assess judges to decide whether to release or detain an arrested person awaiting a trial
Trackur	N/A

Table 24 - Asset assessment for N-T-1

Functionalities / Assets	F1: Technical infrastructure to support new policies and increasing amount of data	F2: Staff training to be able manage and produce "good" data
Aragon Open Data	Availability and reuse of existing public open data, through ontology model and technical infrastructure	Standardised ontology makes it easier for staff and users to expose and reuse data

Functionalities / Assets	F1: Technical infrastructure to support new policies and increasing amount of data	F2: Staff training to be able manage and produce "good" data
Big Data Test Infrastructure (BDTI)	Connecting Europe Facility (CEF) component that provides a set of services to help public administrations explore and experiment with various data sources, software tools and methodologies	Provides support services and documentation to public administrations
Datawrapper	Easy data visualisation	For non-IT staff
FishstatJ	Provides data access to users for specific statistical data about fisheries	For non-IT staff
Galileo	Satellite system that provides increased accuracy in navigation, positioning and timing services	Embedded in smartphones and vehicle navigation systems
MapR	Platform that supports big data management and processing for governments with artificial intelligence and analytics, supporting different scenarios in the public sector competences	Training of staff is required
NodeXL	Tool to explore network graphs created from existing data and even from social network data streams	For advanced users with IT knowledge
OPEN ARTFISH	Smartphone app and database to collect data to know the status and trends of capture of fisheries	For end users
OpenRefine	OpenRefine can help to explore large data sets with ease, cleaning it; transforming it from one format into another; and extending it with web services and external data	For IT staff
SAKE Semantical analysation of complex events	N/A	N/A
Smart Start	Specific programme based on data analysis for supporting beneficial childhood experiences	N/A
Watson Super Computer Project	Supercomputer AI services	N/A

Table 25 - Asset assessment for N-T-3

Functionalities / Assets	F1: Data protection	F2: Information security management
Aragon Open Data	N/A	Open data from different policy domains and departments ready to be used

Functionalities / Assets	F1: Data protection	F2: Information security management
ISO 27001	N/A	ISO/IEC 27001 is the best-known standard in the family providing requirements for an information security management system, helping organizations keep information assets secure
Smart Start	Programme in The Netherlands that develops a data-driven and fact-based approach analysing big data from a wide range of sources to estimate the risk to the child's future well-being	N/A

Table 26 - Asset assessment for N-T-4

Functionalities / Assets	F1:IT infrastructures must be reliable, secure and economically sustainable
Big Data Test Infrastructure (BDTI)	Provides the infrastructure to help public administrations explore and experiment with various data sources, software tools and methodologies.
Blockcerts: An open Standard for Blockchain educational certificates	The citizen can share personal data (blockchain-based certificates about civic records, academic credentials, professional licenses, workforce development, etc.) in a reliable, secure and sustainable way
European Open Science Cloud	It functions as a virtual environment with open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines by federating existing scientific data infrastructures.
Interoperability Centre	Provides a unified infrastructure for the installation and use of online services through which operational data is exchanged between the Ministry of Finance and other public bodies in Greece
Italian Data Analytics Framework (DAF)	Big Data Platform to store in a unique repository the data of the PAs, implementing ingestion procedures to promote standardization and therefore interoperability among them
RapidMiner	Data science solutions
Weka	Machine learning algorithms for data mining tasks

Table 27 - Asset assessment for N-I-1

Functionalities / Assets	F1: Strategic management system integrating both, financial and nonfinancial performance information.
Smart Start	Non-integrated financial and nonfinancial information in this solution
Solver BI360	Tool to make financial and operational reporting form data
The European Data Market Monitoring Tool	N/A
€CONOMIA - The Monetary Policy Game	A simulation game for monetary policy

Table 28 - Asset assessment for N-I-3 - Use Case

Functionalities / Assets	F1: Availability of accurate, accessible, valid, timely complete and relevant information	F2: Organisational conditions established to use the information adequately (employees need to be able to understand and use the information as well as to find creative solutions).
2050 Pathways Web Tool	Tool to collect citizens inputs for decarbonisation in UK towards 2050	Understandable
3D City Model	Open data Adelaide 3D city model to help visualise the City's future, particularly in relation to growth scenarios and land use planning	Easy to use and understand
African Highland Farmer – the Game	Simulation game for policymakers to raise awareness about policies impacts on farmers' decisions and farms production and economic results towards development of domain specific target and indicator systems	Easy to understand and use
Energieatlas Bayern	N/A	N/A
Global Pulse	Promote the use of big data for research and development with a network of innovation labs	Develop toolkits, applications and platforms to improve data-driven decision-making and support evaluation of promising solutions
Google ECO Projects	Information about Google projects and environmental impact	N/A
GovTrack	Provides open information about the US Congress activities	Helps US citizens to participate in their national legislature
In the Air	Visualisation project for microscopic and invisible agents of Madrid´s air	Individual and collective awareness and decision-making support tool
It's Your Parliament	Information on votes cast per members and groups of the European Parliament	Information for public scrutiny
MASAR	Crowd control centre and tracking platform to help visitors plan their routes in Mecca and Medina	Specific control centre
OpenGov.gr	Greece Open government web. Open calls; Electronic deliberation on draft legislation or policy initiatives; Labs for new ideas and proposals from citizens	For citizens and policimakers collaboration on policy making
SeeClickFix	Reporting tool for non-emergencies in communities. Available interface for citizen and for officials	To be deployed by the municipality
Smart City - City Information Modelling Rotterdam	Integration of city information into a 3D model	Allows interoperability among city departments and potential development of new services for city development
Smart Construction Administration	Using sensors to perform maintenance of transport infrastructures, including information from user's smartphones	A collaborative infrastructure has to be set up and integrated with the public administration organisation
Smart Start	Use of big data analytics to achieve a well-being childhood	This must be supported by policy makers and the corresponding programs to achieve the objectives

Functionalities / Assets	F1: Availability of accurate, accessible, valid, timely complete and relevant information	F2: Organisational conditions established to use the information adequately (employees need to be able to understand and use the information as well as to find creative solutions).
X-Road	Estonian e-government platform to provide services to citizens and internally	A new approach to modern e-government

Table 29 - Asset assessment for N-I-3 - Code list / Ontology / Taxonomy / Vocabulary/Standard

Functionalities / Assets	F1: Availability of accurate, accessible, valid, timely complete and relevant information	F2: Organisational conditions established to use the information adequately (employees need to be able to understand and use the information as well as to find creative solutions).
Agrovoc	Multilingual vocabulary for food and agriculture from FAO available as Linked Open Data	Can be browsed on-line, downloaded, and accessed through SPARQL and webservices
DCAT Application Profile for Data Portals in Europe (DCAT-AP)	Description of public sector datasets to enable cross-data portal search for data sets and make public sector data better searchable across borders and sectors	Has to be implemented in public datasets catalogues
FoodEx2	Standardised food classification and description system, facilitating comparison and data analysis	Specific for the food sector and public regulation
OECD Taxonomy of Economic Activities Based on R&D Intensity	Classification of industries according their percentage of R&D investment	Useful for policy makers

Table 30 - Asset assessment for N-I-3 - Application

Functionalities / Assets	F1: Availability of accurate, accessible, valid, timely complete and relevant information	F2: Organisational conditions established to use the information adequately (employees need to be able to understand and use the information as well as to find creative solutions).
ALERTS (Automated Land change Evaluation, Reporting, and Tracking System)	Near real-time land use and land-cover change detection for decision support evaluation, reporting and tracking system	Web based tool
BudgIt	Online information about budget and public finance with different views and levels of detail	Easy to visualise, but more detailed information can be accessed
Buienalarm	Accurate information for rain prediction	N/A
Cool Farm Tool Water	Crops' water needs based on user inputs and global datasets	Information centralised in an application
Diabetes Plus	Diabetes diary to annotate glucose readings, insulin doses and patient activity	Easy tool that can be managed by patients allowing to forward information to doctor

Functionalities / Assets	F1: Availability of accurate, accessible, valid, timely complete and relevant information	F2: Organisational conditions established to use the information adequately (employees need to be able to understand and use the information as well as to find creative solutions).
Electronic Health Records	Improving healthcare delivery through the management of electronic health records for the health public sector	Protects the privacy of the patients
LiquidFeedback	Tool to propose and vote ideas (digital assembly)	Policy-makers can collect inputs from citizens
Meieraha	Estonian budget visualisation	Tool to help citizens understand the country budget
Opinion Crawl	Online sentiment analysis from web	Should be embedded in a wider topic analysis for public administrations
Opinion Space	Tool for the generation and exchange of new ideas about issues and policies	For citizens and policymakers
Runtastic Applications	Exercise and health apps for personal use	N/A
Workday	Enterprise cloud applications for enterprise management	Plan, manage and control organisations with this cloud-based solution
World in figures	Countries profiles and ranking indexes	Can be used as a data source for analysis

Table 31 - Asset assessment for N-I-3 - Tool

Functionalities / Assets	F1: Availability of accurate, accessible, valid, timely complete and relevant information	F2: Organisational conditions established to use the information adequately (employees need to be able to understand and use the information as well as to find creative solutions).
Infogram	Tool offering several charts and maps to visualise data	Useful to help understand data
Italian Data Analytics Framework (DAF)	Italian Government and Public Administration tool to support the diffusion of open data and to enable data-driven policies	Provides public and private access portals
MapR	Platform that supports big data management and processing for governments with artificial intelligence and analytics, supporting different scenarios in the public sector competences	Users require training to understand and use the platform
Open policy making toolkit	Guide for open policy making	For policy makers
Orange	Open source machine learning and data visualization. Interactive data analysis workflows with a large toolbox.	Tool for novice and expert. No need to have programming skills
Qlik	Generic tool that supports big data analytics and AI	To support decision makers with data analytics
Semantria	Cloud sentiment analysis tools, including social media and other sources	Should be embedded in a wider topic analysis for public administrations
Tableau Public	Analytics and visualisation tool	To support decision makers with data analytics

Functionalities / Assets	F1: Availability of accurate, accessible, valid, timely complete and relevant information	F2: Organisational conditions established to use the information adequately (employees need to be able to understand and use the information as well as to find creative solutions).
Virtuose DE	Cloud-based video service platform for the analysis of traffic movements	To be integrated in public traffic control systems

Table 32 - Asset assessment for N-I-3 - Portal/Database/Data source

Functionalities / Assets	F1: Availability of accurate, accessible, valid, timely complete and relevant information	F2: Organisational conditions established to use the information adequately (employees need to be able to understand and use the information as well as to find creative solutions).
Copernicus	Earth satellite and sensor observation information for different domains available through different data services	Must be set up for each need. The Data and Information Access Services can provide tailored services for each need
Employment Ontario Geo Hub	Open data for employment Ontario	Provides the data and analytical tools
ESPON Database for policy makers	Regional indicators database	Focused towards Policy Makers, as well as towards scientists too
EU Open Data Portal	Provides access to an expanding range of data from the European Union (EU) institutions and other EU bodies	Data is organised by categories
EU Science Hub	Compilation of open databases and tools from projects	Organised by research area
EUMETSAT	Satellite climate and environmental data	Provides several data services
European Data Portal	Metadata catalogue from Public Sector Information data	Data is organised by categories
Europeana	Open database with digitised cultural contents from European archives, libraries and museums	Organised by collections, exhibitions and exploration tools
Galileo	Satellite system that provides increased accuracy in navigation, positioning and timing services	Embedded in smartphones and vehicle navigation systems
RASFF Database	Tool to get alerts on food. Publicly available	Easy to use, even for citizens
The CIARD Routemap to Information Nodes and Gateways (RING)	Directory of datasets and data services for agri-food sector	Mainly for agricultural information professionals and data scientists.

Table 33 - Asset assessment for N-I-3 - Model

Functionalities / Assets	F1: Availability of accurate, accessible, valid, timely complete and relevant information	F2: Organisational conditions established to use the information adequately (employees need to be able to understand and use the information as well as to find creative solutions).
Economic Simulation Library	N/A	N/A
GLEAM	Tool for simulation of human mobility and disease transmission based on real data	Simulation tool for public authorities and policy makers

Table 34 - Asset assessment for N-I-4

Functionalities / Assets	F1:Knowledge in the Public Sector should be collected, stored, shared and eventually destroyed.
Big Data Test Infrastructure (BDTI)	Provides the infrastructure to help public administrations explore and experiment with various data sources, software tools and methodologies.
Digital Policy Model Canvas	Methodology that can help guide policymakers. It is a canvas approach that helps translate broad insights and understandings to the needs of a particular country. It also helps define the key issues at stake as well as metrics to evaluate success, and suggest avenues for possible iteration and improvement
European Data Portal	This portal harvests the metadata of Public Sector Information available on public data portals across European countries
OPEN ARTFISH	Toolkit for routine small-scale fisheries data collection. Its objective is to facilitate the implementation of cost-effective and sustainable routine data collection, storage and analysis of data, using the appropriate statistical procedures
Open policy making toolkit	Contains the tools and techniques needed to run through diagnosis, discovery and idea generation
OpenAIRE	Shifts scholarly communication towards openness and transparency and facilitate innovative ways to communicate and monitor research.
Qlik	Public sector organizations have tremendous amounts of siloed data. By combining all these data and making it easy for everyone to explore, Qlik delivers the valuable insights needed to efficiently improve services
Semantria	Business intelligence solution focused on drawing insights from unstructured text data
Smart Start	Programme in The Netherlands that develops a data-driven and fact-based approach analysing big data from a wide range of sources to estimate the risk to the child's future well-being
SmartRegio	Management Consultant for Smart Energy in rural regions. Provides statistics from social media platforms as well as individual data of little regions in terms of mobility, energy and so on
Tableau Public	It provides with speed, accuracy, transparency and ease of communication to the Government analytics
The OO Software	System that helps with backup of information and recovery
X-Road	A platform that allows the secure exchange of data in order to provide efficient public services. The tool can write to multiple databases, transmit large data sets and perform searches across several databases simultaneously. It gives a seamless service provision for citizens, given that once the data is updated, all other service providers will automatically also operate with up to date information