Roadmap for Future Research Directions Research Challenges

WORK IN PROGRESS

Please provide feedback to francesco.mureddu@lisboncouncil.net and https://roadmap.bigpolicycanvas.eu/

Document Information

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

Table of Contents

List of Tables

List of Figures

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

1.1 Research Clusters

We define five 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 the fifth one is transversal at each phase of the cycle (Figure 1).

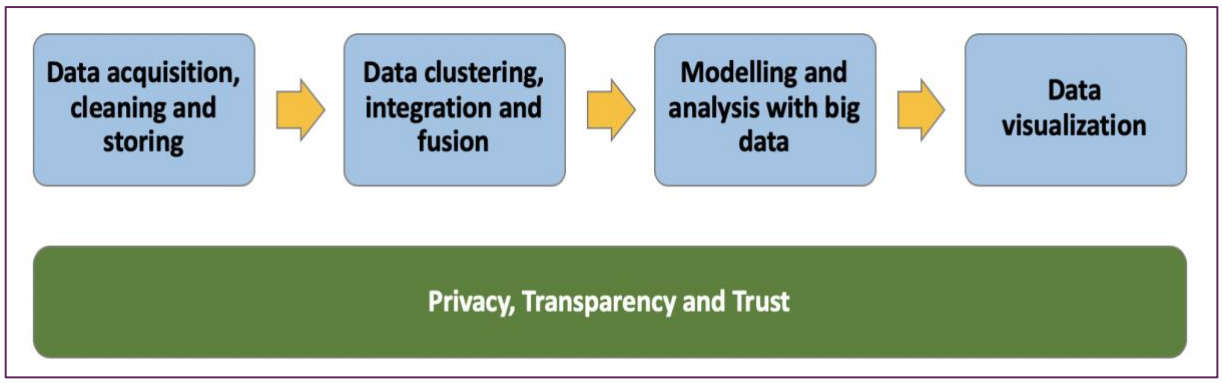

Figure 1 – Structure of the Research Clusters

Let us present know each cluster more thoroughly.

1.1.1 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[1], or the use of algorithms for the automatic detection of fraudulent behaviour 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. 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.

1.1.2 Cluster 2 - Data Acquisition, Cleaning and Storing

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

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

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. 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. 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. 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. 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.

1.1.3 Cluster 3 - Data Clustering, Integration and Fusion

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 an 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. 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.

1.1.4 Cluster 4 - 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.

1.1.5 Cluster 5 - 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.[2] 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 create 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. 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 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

[2] 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

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.

1.2 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 1.

Table 1 – Research clusters and related research challenges

Research Cluster	Research Challenges
C1- Privacy, Transparency and Trust	RC 1.1 - Big Data nudging
	RC 1.2 - Pervasive data collection
	RC 1.3 - Algorithmic bias and transparency
	RC 1.4 – Open Government Datasets
C2 - Data acquisition, cleaning and representativeness	RC 2.1 – Real time big data collection and production
	RC 2.2 - Quality assessment, data cleaning and formatting
	RC 2.3 - Representativeness of data collected
C3 - Data clustering, integration and fusion	RC 3.1 - Identification of patterns, trends and relevant observables
	RC 3.2 - Extraction of relevant information and feature extraction
	RC 3.3 - Integration and interoperability of Public Administration datasets
C4 - Modelling and analysis with big data	RC 4.1 - Identification of suitable modelling schemes inferred from existing data
	RC 4.2 - Collaborative model simulations and scenarios generation
	RC 4.3 - Integration and re-use of modelling schemes
	RC 5.1 – Automated visualization of dynamic data in real time

C5 - Data visualization	RC 5.2 - Interactive data visualization

1.2.1 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 channelling 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.

Importance 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. 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 and tools. 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.

RC 1.2 - Pervasive data collection

Description. It is well known that the amount of data produced is increasing exponentially, government data are no exception. Massive interconnection, i.e. massive number of objects/things/sensors/devices connected 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 stretch the applicability of traditional safeguards such as informed consent and anonymization. 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) by adopting a risk-based approach. 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.

Importance 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 and once only principle only increase the strategic importance of personal data, but simultaneously highlight the urgency of identifying workable solutions.

Technologies and tools. Several tools are today being developed in this area. 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).

RC 1.3 – Algorithmic Bias and Transparency

Description. Many decisions, in the public as well as in the private sector, are today automated and performed by algorithms. Predictive algorithms have been used for many 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.

Importance in policy making process. 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.

Technologies and tools. 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: awareness, access and redress, accountability, explanation, data provenance, auditability, validation and testing.[3] Further, Geoff Mulgan from NESTA has developed a set of guidelines according to which governments can better keep up with fast-changing industries.[4] Similarly, Eddie Copeland from NESTA has developed a "Code of Standards for Public Sector Algorithmic Decision Making."[5]

RC 1.4 – Open Government Data

Description. 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.[6] 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.

Importance in policy making process. 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.

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

[4] 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/

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

[6] For more information please refer to https://www.finance.gov.au/sites/default/files/Big-Data-Strategy.pdf

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 and tools. 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. Inspiring cases include: Open Government Initiative[7] carried out by the Obama Administration for promoting government transparency on a global scale; Data.gov:[8] 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[9] (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:[10] 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:[11] 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: [12] website providing a window of US Government Administration effort to improve performance and accountability.

1.2.2 Research Challenges on Data acquisition, Cleaning and Representativeness

RC 2.1 - Real time big data collection from networks

Description. 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

[7] www.whitehouse.gov/open
[8] www.data.gov/
[9] www.opengovplatform.org/
[10] www.usaspending.gov/
[11] For more information please refer to www.federalregister.gov/
[12] For more information please refer to www.performance.gov/

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).

Importance 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.

Technologies and tools. For collecting the data from devices, an obvious choice is given by the Internet of Things technologies. 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.

RC 2.2 - Quality assessment, data cleaning and formatting

Description. 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. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting. 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.

Importance in policy making. Apart from systematic errors in data collection, it is important to assess to extent 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.

Technologies and tools. 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 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). The Big Data Quality framework developed provides a structured view of quality at three phases of the business process: i.e. Input (acquisition analysis of the data); Throughput (transformation, manipulation and analysis of the data); Output (the reporting of quality with statistical outputs derived from big data sources). Likewise, Taleb et al. (2018) provide a Big Data Quality Assessment Model for Unstructured Data,

RC 2.3 – Representativeness and validity of data collected

Description. 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. As explained by Buelens et al. (2014): "A subset of a finite population is said to be representative of that population with respect to some variable, if the distribution of that variable within the subset is the same as in the population. A subset that is not representative is referred to as selective." 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. 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). Other challenges, identified by Tufekci (2014), include sampling biases arising from selection by hashtags, vague and unrepresentative sampling frames, sociocultural complexity of user behaviour aimed at algorithmic invisibility (such as sub-tweeting, mock-retweeting, use of "screen captures" for text), as well as accounting for field effects, i.e. broadly consequential events that do not diffuse only through the network under study but affect the whole society.

Importance in policy making process. Clearly big data representativeness is crucial to policy making, especially when studying certain characteristics of the population and in analysing its sentiment. 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 and tools. 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. The first method uses a version of inverse sampling by incorporating auxiliary information from external sources, and the second one borrows the idea of data integration by combining the big data sample with an independent probability 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.

1.2.3 Research Challenges on Data Clustering, Integration and Fusion

RC 3.1 - Identification of patterns, trends and relevant observables

Description. 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 have not been previously visible.

Importance in the policy making process. Clearly the possibility to extract patterns and trends in data can help the policy maker in having a first sight for discovering issues that are the used to develop the policy agenda. An interesting application is anomaly detection, which is most 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 and tools. 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 in order to assess some employee characteristics or consumer bundle analysis to model the behavior of customers. It has also to be taken into account that most of the Big Data are not structured and have a huge quantity of text. In this regard, text mining is another technique that can be adopted to identify trends and patterns.

RC 3.2 - Extraction of relevant information and feature extraction

Description. Summarizing data and meaning extraction to provide a near real time analysis of the data. Data must be structured prior to the analysis in an homogeneous way, as algorithms unlike humans are not able to grasp nuance. 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. Even after cleaning and error correction in the database, some errors and incompleteness will remain, challenging the precision of the analysis.

Importance in policy making process. 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, ensuring the widest reuse of the data for purposes different from the one it was originally gathered. The data have to be adapted according to the use and analysis that are destined too, and moreover they are needed as data preparation for visualization;

Technologies and tools. Bayesian techniques for meaning extraction; extraction and integration of knowledge from massive, complex, multi-modal, or dynamic data; data mining; scalable machine learning.

RC 3.3 - Integration and interoperability of PA datasets

Description. The integration and interoperability of government data is a long standing issue, but it 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. At the same time, the issues of data centralization versus federation, as well as data protection, remain challenges to be dealt with. 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.

Importance in policy making process**.** 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 and tools. New interface 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. As an example, the Data & Analytics Framework (DAF) by the Italian Digital Team aims to develop and simplify the interoperability of public data between PAs, standardize and promote the dissemination of open data, optimize data analysis processes and generate knowledge. In this regard, the benefits for public administrations will be the following:[13]: significantly enhance the value of PA's information assets through the preparation and use of analytical tools designed to synthesize knowledge for decision makers, and the dissemination of information to citizens and businesses; optimize data exchange between PAs and Open Data deployment, minimizing transaction costs for data access and usage; facilitate data analysis and data management by data scientist teams within the PA, in order to improve knowledge of the phenomena described by the data and develop "intelligent" applications, as well as take initiatives to promote scientific research activities on application themes of interest to the PA. Another interesting example is given by the X-Road, which 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.[14]

[13] For more information please refer to https://teamdigitale.governo.it/en/projects/daf.htm
[14] For more information please refer to https://e-estonia.com/solutions/interoperability-services/x-road/

1.2.4 Research Challenges on Modelling and Analysis with Big Data

RC 4.1 - Identification and validation of modelling schemes inferred from existing data

Description. The traditional way of modelling started with a hypothesis about how a system acts. Then collect data to represent the stimulus. 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. In this respect, the real challenge is to be able to identify and validate from existing data models that are valid and suitable to cope with complexity and unanticipated knowledge. Model 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 and that the model's representation of the problem entity and the model's structure, logic, and mathematical and causal relationships are "reasonable" for the intended purpose of the model. A second phase is the computerised model verification ensures that computer programming and implementation of the conceptual model are correct, as well as states that the overall behaviour of the model is in line with the available historical data. Finally, model validation is also very important when models need to be re-used (more on that on challenge 4.3).

Importance in policy making process. There are several aspects related to the identification 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 and tools. 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. Validation and verification process of a model is never completed.

RC 4.2 – Collaborative model simulations and scenarios generation

Description. 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. 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. Refers to a process where a number of people actively contribute to the creation of a model. 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 (decision makers, analysts, companies, civic society, and the general public.
Importance in policy making process. 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 and tools. The Citychrone++ platform[15] provides users with a powerful tool to assess the present status of urban accessibility and design and evaluate future scenarios. It integrates flexible data analysis tools with a simple scenario simulation platform in the area of urban accessibility. In this framework, it will be important to parallel the platform with effective modelling schemes, key for the generation and the assessment of new scenarios. CityChrone++ 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. CityChrone++ focuses specifically on urban 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.

RC 4.3 - Integration and re-use of modelling schemes.

Description. 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

[15] http://whatif.cslparis.com/citychrone.html

and methodologies by using existing ones or by defining new ones. Further to that, the present sub-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.

Importance in policy making process. In systems analysis, it is common to deal with the complexity of an entire system by considering it to consist of interrelated sub-systems. 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.

Technologies and tools. 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.

1.2.5 Research Challenges on Data Visualization

RC 5.1 - Visualization of dynamic data in real time

Description. 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.

Importance in policy making process. 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 and tools. 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.

RC 5.2 - Interactive visualization of simulation results

Description. 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.

Importance in policy making process.

Policy makers should be able to independently visualize results of analysis. In this respect, benefits of interactive data visualization are basically of three main typologies. The first, is 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. Secondly, it helps to understand the impact of policies: interactive visualization is instrumental in making evaluation of policy impact more effective.

Technologies and tools. 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. 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.

References

- Buelens, B.et al. (2014). Selectivity of Big Data. Discussion paper, 2014/11, Statistics Netherlands, The Hague/Heerlen, The Netherlands.
- 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*.
- Kim, J.K., Berg, E. & Park, T. (2016). Statistical matching using fractional imputation, Surv. Methodol., 42, 19–40.
- 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.
- Kim, Jae Kwang and Wang, Zhonglei, "Sampling techniques for big data analysis in finite population inference" (2018). Statistics Preprints. 136.
- 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.
- 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.
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*.
- 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.
- 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.
- Robinson, D., & Yu, H. (2012). The New Ambiguity of Open Government. UCLA Law Review Discourse, 1-17.
- Ruths, D., & Jurgen, P. (2014). Social media for large studies of behavior. *Science*, 1063-1064.
- Selva Rathna and T. Karthikeyan, "Survey on Recent Algorithms for Privacy Preserving Data mining", IJCSIT, Vol. 6 (2) , 2015, 1835-1840, ISSN: 0975-9646.
- Social Media Research Group (2016). *Using social media for social research: An introduction*. UK: Government social research profession
- Taleb, Ikbal & Serhani, Mohamed & Dssouli, Rachida. (2018). Big Data Quality Assessment Model for Unstructured Data.
- The Centre for Public Impact (2017). Destination unknown: Exploring the impact of Artificial Intelligence on Government
- 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.
- UNECE (2014). A Suggested Framework for the Quality of Big Data. UNECE, December 2014.
- 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.