ERCIM Expert Group on "Big Data Analytics"

Big Data Analytics: towards a European research agenda

Editors: Mirco Nanni[1], Costantino Thanos[1], Fosca Giannotti[1] and Andreas Rauber[2]

Executive Summary

Big data are blossoming together with the hope to harness the knowledge they hide to solve the key problems of society, business and science. However, turning an ocean of messy data into knowledge and wisdom is an extremely challenging task.

In this paper we put forward our vision of Big Data Analytics in Europe, based on the fair use of big data with the development of associated policies and standards, as well as on empowering citizens, whose digital traces are recorded in the data. The first step towards such objective is the creation of a European ecosystem for Big Data Analytics-as-a-service, based on a Federated Trusted Open Analytical Platform for Knowledge Acceleration. The goal is to yield a data and knowledge infrastructure providing to citizens, scientists, institutions and businesses: (i) access to data and knowledge services, (ii) access to analytical services and results, within a framework of policies for access and sharing based on the values of privacy, trust, individual empowerment and public good. Several requirements need to be fulfilled, at least at four very different levels (the four dimensions discussed in detail in this paper):

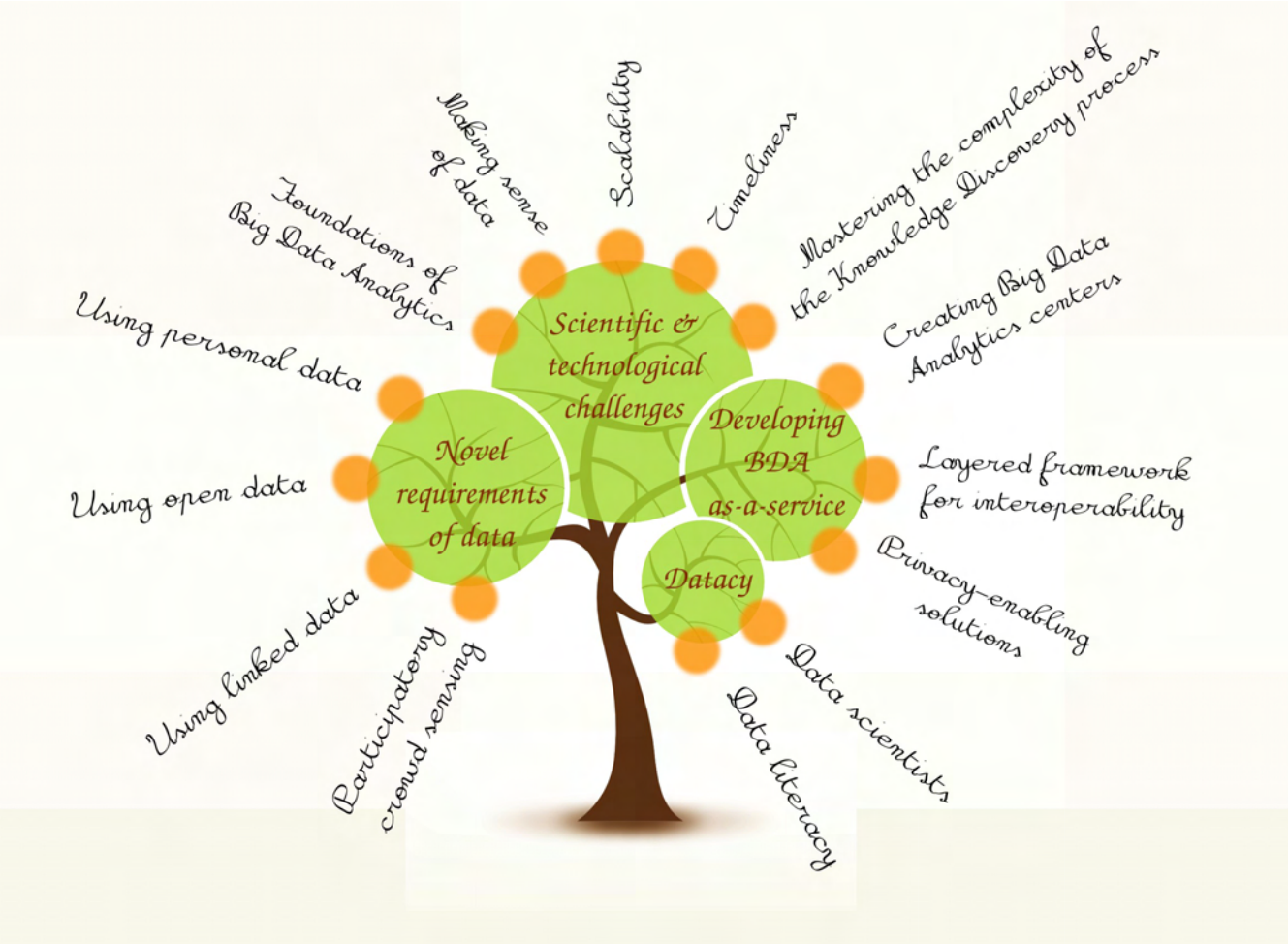
- **Scientific and technological challenges**. Solutions to difficult problems are needed, including**:** i) the development of new foundations for Big Data Analytics, which integrate knowledge discovery from Big Data with statistical modeling and complex systems science, ii) semantics data integration and enrichment technologies, which make sense of Big Data and make them usable for high-level services, iii) scalable, distributed, streaming Big Data Analytics technologies, which master the scary volume and speed of Big Data.
- **Data requirements**. The potential value that lies in Big Data can be unleashed only if a proper, efficient, fair and ethical access to data is provided to the relevant actors. This poses many techno-social questions: who owns and use personal data? What is the real value of such data? How to make it possible to access and link the different data sources? How to empower individuals by providing the capabilities for accessing, using, handling and controlling the usage of own data? How to have individuals and institutions understand the social and/or economic impact of personal data? How to boost open data initiatives and the development of federations of linked data?
- **Education and data literacy**. Realizing wisdom and value from big data requires competent professionals on data analytics having, among others, deep expertise in statistics and machine learning. Skills must be developed on how to exploit data and their analysis to develop successful business initiatives. Moreover, given the pervasiveness of Big Data in most of the disciplines of human knowledge and research, elements of data science should be provided to students of all levels of education, from high-schools to university curricula.
- **Promotional initiatives for data analytics and BDA-as-a-service**. Letting BDA-as-a-service have a chance to flourish in Europe requires not only technical advancement but also several

1 Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", CNR, Italy. {name.surname}@isti.cnr.it
2 Department of Software Technology and Interactive Systems, TU-Wien, Austria. rauber@ifs.tuwien.ac.at

organizational actions aimed to promote development along some key directions, such as supporting the creation of Big Data Analytics centers accessible to researchers, public administrations, medium and small companies; incentivize the adoption of a layered framework to increase interoperability across the single data repositories; aim to reach a European leadership in the development of privacy-enabling solutions.

Finally, a most effective way for promoting and helping the development of Big Data Analytics is to create successful, large-scale showcases in high-impact application domains. Among the several possible ones, we recommend to explore the context of Smart cities and communities, Big Data Analytics for developing countries, the management of the global market of jobs, the (quantitative) assessment of results of European projects and activities, and the development of Big Data-aware Official Statistics.

The final recommendations we provide to EU as conclusion of this paper can be summarized in the investment over the four dimensions listed above (technologies, methods and norms for data access, education, promotion of BDA-as-a-service), in particular by spawning federations of key public and private actors, in challenging multidisciplinary domains to provide a critical mass for starting up Federated Trusted Open Analytical Platforms for Knowledge Acceleration and creating incentives for further actors to join



Visual representation of the four dimensions of Big Data Analytics and their main topics

(Template image from http://dryicons.com)

1 The Big Data Landscape

According to the forecast of a recent report [4], "if the right investments and decisions are made in the coming years, Europe can leverage the fast growing market of Big Data so that by 2020 it will play a leading role in the global market around the creation of value from Big Data. […] By the end of this decade, data business has become a key industry in Europe developing products and services around data itself, the analysis of data, and by using the insights gained by analysing data. Data-driven applications will help companies to design better products, to improve their business plans, and to create new business models. They will help governments to implement policies more effectively and individuals to improve the quality of their lives. People will trust those data-driven applications and will use them broadly." In the scientific disciplines, a new data-dominated science is emerging, which moves towards a data-centric way of thinking, organizing, and carrying out research activities that could lead to a rethinking of new approaches to solve problems that were previously considered extremely hard or, in some cases, even impossible to solve and also lead to serendipitous discoveries.

Put another way, big data are blossoming together with the hope to harness the knowledge they hide to solve the key problems of society, business and science. However, turning an ocean of messy data into knowledge and wisdom is an extremely difficult task. "We are drowning in data and starving for knowledge"[3].

The amount of data in our world has been exploding. *Science* gathers data at an ever-increasing rate across all scales and complexities of natural phenomena. New high-throughput scientific instruments, telescopes, satellites, accelerators, supercomputers, sensor networks and running simulations are generating massive amounts of scientific data. *Companies* capture trillions of bytes of information about their customers, suppliers, and operations. *Smart sensing*, including environment sensing, emergency sensing, people-centric sensing, smart health care, and new paradigms for communications, including email, mobile phone, social networks, blogs, Voip, are creating and communicating huge volumes of data.

By Big data we mean datasets whose size and uses are beyond the ability of traditional database software tools to capture, store, manage, and analyze. But the science of Big Data is not just about volumes and velocity of data, but it also deals with: heterogeneity (levels of granularity, media formats, scientific disciplines involved), and complexity (uncertainty, incompleteness, representation types).

More business and government agencies are discovering the strategic uses of large data collections. And as all these systems begin to interconnect with each other and as powerful new software tools and techniques are invented to analyse the data for valuable inferences, a radically new kind of "knowledge infrastructure" is materializing. A new era of Big Data is emerging, and the implications for business, science, government, democracy and culture are enormous.

The inferential techniques being used on Big Data can offer great insight into many complicated issues, in many instances with remarkable accuracy and timeliness. The quality of business decision-making, government administration, scientific research and much else can potentially be improved by analysing data in better ways.

Big Data technologies have the capability to create an ecosystem of novel data-driven business opportunities, facilitated by participatory platforms, that could help position Europe's companies to collaborate on uncovering new local, national and global whitespace markets and unleash a new wave of business opportunities based on the concept of innovation from information, leveraged by a novel way for collaborative, participatory creation and enrichment of Big Data.

Big Data presents many exciting opportunities to improve modern society and boost social progress:

3 Quote attributed to Rutherford D. Rogers, a librarian at Yale, by the New York Times in 1985.

support policy making, novel ways of producing high-quality and high-precision statistical information, to empower citizens with self-awareness tools, and by promoting ethical uses of big data. Data science may empower citizens, NGOs and policy makers with the means to gain a better understanding of complex socio-economic systems, methods for introspection of complex global processes, tools for assessing the implications of decisions beforehand, and hence to improve our capacity to sustainably manage our society on the basis of well-founded knowledge and inclusive participation. In particular, data science may provide policy makers with a much deeper understanding of behaviour and interactions between global systems and will enable tools to develop and trial policy in silico.

The objective of this document is to identify the research challenges that should be at the center of the European agenda on Big Data Analytics for the next years. The distinctive aim of this paper is to bring together the technological issues of Big Data management with the analytical challenges of extracting knowledge from Big Data as we believe that the management and the use of data strongly depend on each other. From one side, Big Data Analytics is often about secondary, unforeseen usage of data originally acquired for different purpose. On the other side the truth of the discovered knowledge depends on the quality of the overall data management process of complex data sources. This is a complex process involving a mix of methodologies coming from different research communities, such as statistics, machine learning, data mining, visual analytics, etc., and is aimed to a broad range of application tasks, including data summarization, classification & prediction, correlation analysis, etc.

Context

Data Analytics (without the “Big” adjective) refers to the discovery of useful patterns in data and their effective communication to the user, for instance through appropriate (possibly interactive) visualization techniques. This process usually employs a mix of methodologies coming from different research communities, such as statistics, machine learning, data mining, visual analytics, etc., and is aimed to a broad range of application tasks, including data summarization, classification & prediction, correlation analysis, etc.

In literature, different aspects of what Big Data means have been mentioned within different communities, according to their topics of interest and priorities. Recently some kind of consensus has been reached, listing a set of attributes (all purposely having the same initial letter, forming four Vs) that characterize Big Data most: velocity, volume, variety and veracity. Velocity relates to the streaming nature of data and its speed; volume is about the sheer size of such data, calling for filtering/compression strategies or special measure for storing and processing it; variety stands for heterogeneity of the data sources that can be involved, both in terms of formats and representation, and in terms of semantics; veracity points to data quality issues and trustworthiness of the information. In recent literature a fifth V has been added to the list, standing for value, aimed to emphasize the fact that turning Big Data sources and associated analytics tools into value (economic, social or scientific) is both important and far from trivial. Finally, an additional feature, emphasized by the distributed computing community, is the fact that Big Data usually is distributed in nature, and therefore the issues related to information transportation from source to storage/analysis sites need to be considered.

We can distinguish two large areas that involve Big Data: scientific data and social data (at large). The first area includes data obtained through scientific experiments and/or observations, typically grant some control over the way data are generated and are accompanied by meaningful metadata that specify characteristics and semantics of the collected data. The second area mainly includes data obtained as byproduct of human interactions with ICT services, such as mobility traces, economic transactions, interactions on social media, etc., typically characterized by a lack of control over data generation and poor availability of semantics. A foremost issue associated with data of the latter kind is the possible emergence of ethical issues, such as the risk of harming individual privacy whenever

personal information is involved in the analyses or the possible discriminative use of the knowledge extracted from such data [11]. That led several researchers and public bodies to discuss the actual privacy risks in BDA, as well as to better understand the balance between risks and opportunities. Examples of this include the “90-day review of big data and privacy” recently called by the U.S.A. President [27], World Economic Forum reports on the subject [26], various overviews and working groups [5 and 21].

2 Vision: towards Big Data Analytics-as-a-service

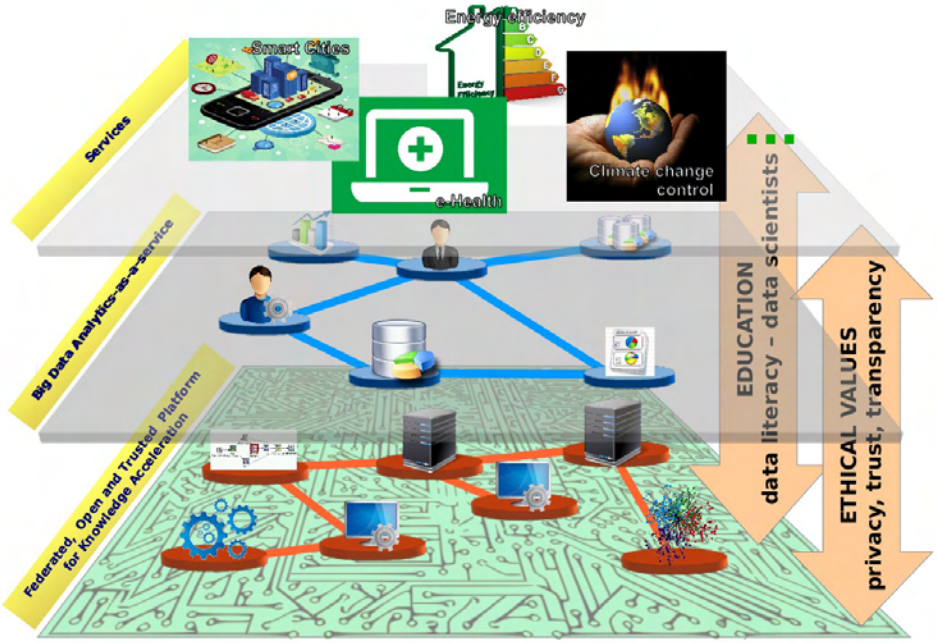
The general, far-fetched vision proposed in this document is based on the fair use of big data with the development of associated policies and standards as well as empowering those whose digital traces are recorded in the data, thus providing a conceptual framework for exploring the past, present and near future(s) by making sense of the digital traces – a sort of *digital time machine*. In the first place, the digital time machine is meant to empower individual citizens, that can explore the past and the present to gain better knowledge of self and own position in society, and explore plausible futures to reason on the consequences of decision making. But the digital time machine also works for communities, institutions, and businesses. It is based on a fair use of big data, rooted on ethical values such as trust, privacy, transparency, and public good.

In our view, a first step towards the realization of this framework is to achieve an extremely important milestone that should be put in the EU research agenda of the next 10 years: the creation of an European ecosystem for **Big Data Analytics-as-a-service** based on a Federated Trusted Open Analytical Platform for Knowledge Acceleration. The goal is to yield a data and knowledge infrastructure providing to citizens, institutions and businesses: (i) access to data and knowledge services, (ii) access to analytical services and results, within a framework of policies for access and sharing based on the values of privacy, trust, individual empowerment and public good. In order to achieve this, several requirements need to be fulfilled, at very different levels:

- Provide ***norms*** (to respect privacy, manage trust, empower individuals)***, policies, standards, etc***. to access, share, curate, manage and analyze Big Data. These should cover both aspects relative to intellectual property rights (for instance to grant guarantees to the providers of precious datasets in a scientific data center), privacy protection (as in the case of social data) and other ethical issues related to the usage of knowledge, such as discrimination. At the present, the big data collectors as YGAFA (Yahoo-Google-Apple-Facebook-Amazon) and TELCO operators (collecting mobile phone records) have complete control over the individual data they gather, and nobody else has any real vision or decisional power about such informations. Our view is that this status should change along at least two directions. On one hand, the individuals should be granted the right to control and decide about the personal data regarding themselves, including the right to retract or trade them. On the other hand, society at large, including public administration, single individuals, and private companies, should benefit from these big data, through a fair and regulated access to data and/or analytical services.
- Provide the appropriate ***technology for data access*** in order to allow users to efficiently localize the information relevant to specific analytical goals, retrieve all data needed, link and seamlessly connect data coming different sources, and empower the individual in managing his/her personal data – possibly spread across different sites. On one hand, this can be seen as the technological premise for enabling the data accessibility mentioned above. On the other hand, even the basic data management functionalities need to be adapted and revised in order to cope with the many issues introduced by big data, such as: the extreme growth rate of data volumes, that is going to make permanent storage unsustainable; the several heterogeneous data sources that need to be integrated, that makes traditional data modeling and integration procedures inapplicable at a large scale; ensuring the trustworthiness of the data and of the process flows applied on them.

- Provide the appropriate ***analytical technology*** for distributed data mining and machine learning for big data, and a solid statistical framework adapting standard statistical data generation and analysis models to big data: once again, the sheer size and the complexity of big data call for novel analytical methods. At the same time, the kind of measures provided by the data and the population sample they describe cannot be easily modeled through standard statistical frameworks, which therefore need to be extended to capture the way the data are generated and collected.
- Provide ***training programs*** to develop a new generation of data scientists, the emergent and most wanted professional figure combining "the skills of software programmer, statistician and storyteller/artist to extract the nuggets of gold hidden under mountains of data" [8], to provide scholars from every area with a background on data analytics, and to promote *datacy* – i.e., basic literacy on data, their potential value and their risks.

These general requirements lead to define several challenges for the scientific research, as well as for the societal growth and well-being, which are discussed in the remaining of this paper.



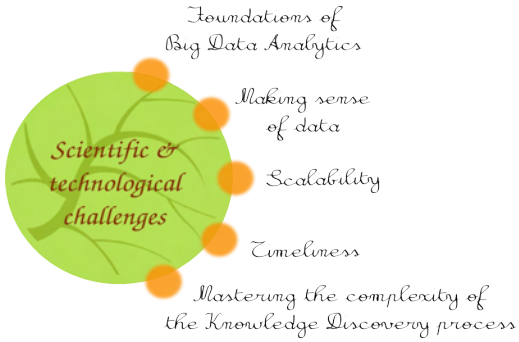
The vision: an ecosystem of users, competences, analytic functionalities, etc. for Big Data Analytics-as-a-service, grounded on a Federated Trusted Open Analytical Platform for Knowledge Acceleration, and basis for Big Data applications and services.

3 The four dimensions

In this section we discuss four different classes of challenges that have to be faced to provide the basis for the European ecosystem for Big Data Analytics-as-a-service discussed above. Such challenges include (i) the solution of some major technical issues, (ii) the definition of policies and norms to allow proper data access, sharing and management, (iii) the development of the novel cultural background for data scientists, professionals and simple citizens, and (iv) the promotion of actions aimed to make Big Data Analytics-as-a-service more likely to prosper in Europe.

3.1 Scientific and technological challenges

The core requirement for the proposed vision of Big Data Analytics is to develop solutions to the several technical issues that emerge in handling Big Data.



■ Foundations of Big Data Analytics

The common view today about Big Data is that "even noisy Big Data could be more valuable than tiny samples because general statistics obtained from frequent patterns and correlation analysis usually overpower individual fluctuations and often disclose more reliable hidden patterns and knowledge" [1] or, in other words, the sheer size of Big Data is enough to compensate for any bias or defect they might contain. We strongly believe this is an oversimplification of a complex issue. A first important aspect to face is the need of a strong statistical basis to **assess the quality of data** – an aspect that, opposed to more traditional settings, is usually not under the control of the analyst and can change in time, thus requiring a continuous **monitoring of data quality** –, the reliability of the inferences performed over Big Data and to develop quality-aware methods. Indeed, the processes that underlie the generation and collection of Big Data might violate the assumptions generally accepted in standard statistical modeling, and therefore the existing statistical frameworks need to be extended to better fit the new context. In summary, we need a conceptual framework **reconciling Big Data Analytics and statistical modeling**. A specific direction on this topic is the creation of models for synthetic generation of realistic data and null models for benchmarking, i.e. **Big Data generators**.

A second aspect to consider is the combination of models and methods of different kind that can complement each other and therefore better exploit the variety and complexity of information usually carried by Big Data. This combination can apply to several levels: (i) **mixed macro- and micro-models**, i.e. combining global analytical models with local patterns; (ii) a combination of **personal vs. collective profiling, patterns and indicators**; (iii) the integration of mathematical models with **Big Data and social simulation** for realistic what-if reasoning. These combinations of solutions also exemplify the ongoing process of interdisciplinary interaction among scientific communities that is apparently indispensable to face all the complexity of the world described by Big Data. This means the **convergence of several disciplines and technologies**, such as: Data Mining, Machine Learning, Information Retrieval, Natural Language Processing, Statistics, Applied Mathematics, Complex Systems.

Finally, very often Big Data are dynamic and streaming, continuously describing phenomena that change with time. Beside considering scalability issues (postponed to later parts of this section), analyzing this kind of data might require to develop analytical models that are natively time-aware, i.e. time should be part of the model, for instance to perform temporal analysis of spreading phenomena, and that take into consideration **concept drifting** or, more generally, be sensible to sudden changes in the data and quickly reflect them into the models built.

■ Making sense of data

A fundamental requirement for a successful data analytics is to have access to **semantically-rich data** that connect together all the relevant pieces of information for a given analytical objective. In contrast, the common situation with Big Data is that information come from several different data sources of different nature, form (e.g. structured vs. unstructured) and quality level, resulting in

heterogeneous data flows that are difficult to integrate. Also, sensed data are often low-level and semantically poor, since they simply expose the raw details of the measurements enabled by the ICT infrastructure that generate them. Novel methods for heterogeneous data sources as well as data fusion and aggregation methods able to combine the different sources are needed, also capable of dealing with the integration of hundreds or thousands of datasets – a scalability issue deemed to be one of the key challenges for the next future [5] – as well as properly adopt machine learning and data mining methods to infer the semantics hidden in data.

A very important example in this direction is the **extreme integration of personal digital breadcrumbs**: a single individual can generate several kinds of traces (purchase transactions, GPS mobility traces, posts on social networks, etc.) that are collected by different systems, each having its objectives and adopting its formats and policies. Reconstructing his/her whole activity requires to link, integrate and align all the relevant pieces of information to form a personal data warehouse of the individual (we postpone any discussion about the obvious concerns on privacy to the next sections) where the different aspects of the personal activities are represented as axes of a **multidimensional view** of the individual, from which in-depth analyses can start.

From a different viewpoint, the complexity of Big Data also comes from the availability of **complex data types**, each requiring specific analysis tools. In several cases, such as network and graph information, larger amounts of data also lead to have more complex structures to deal with, for instance huge, highly connected and highly multi-dimensional social networks, with several nodes but also several different kinds of links among them. This leads inevitably to scalability issues that go beyond plain management of large data volumes, and call for ad hoc solutions for each specific domain. Similar arguments can be formulated for other data types of great and growing importance, that include text information (web pages, messages of any kind exchanged among users, digitized documents, etc.), sensor data (timeseries of environment measures, traffic flows on the roads, etc.), multimedia (including videos, music, speech, possibly linked to text and sensor data), etc.

As a significant part of Big Data are actually user-generated data collected through social media, dealing with textual information expressed in natural language is a mandatory commitment. In the multi-cultural and linguistically rich environment of Europe (European Union recognises 20 official languages and about 60 other minor ones, and adopted in November 2005 the first Commission

Communication that explores the area of multilingualism), that invariably means also to deal with multi-lingual data sources.

■ Scalability

By definition Big Data imply the treatment of huge amounts of data whose size and features clearly challenge the traditional computational paradigms, calling for suitable solutions in terms of **scalable architectures and algorithms**. The point is, data volume is growing faster than the sheer computational power, and therefore the analysis of data needs necessarily to be approached from different angles. A first example is the redefinition of **Machine Learning** tasks through a distributed paradigm. Here, distribution can help in two different ways: clearly, the overall computation load can be divided over several nodes, aiming to solve difficult problems through the combination of medium-power computational units; in addition, data often comes distributed by nature, therefore computing through distributed schemes can mean to push computation closer to the data sources, naturally avoiding (or mitigating) huge issues in data transfer (raw data needs not to move massively across the nodes of the system) and storage (most raw data are processed on the fly and might be discarded, unless data persistence is strictly required).

Scalability emerges as killer issue also in some areas where traditionally performances (though

always relevant) are not the primary goal. One example is the construction of **Statistical Models** from data, which often requires very expensive processes mainly conceived having in mind the framework and data scale of traditional statistics, where data quality is high, yet its size is relatively modest. The size of Big Data make such processes not affordable anymore, and thus require a thorough revision (also motivated, as discussed above, by a mismatch between typical assumptions made in statistical modeling and actual features of Big Data). Similarly, **Visual Analytics** traditionally focuses on rendering insights of the information hidden in the data and its semantics through effective visual metaphors. The explosion in the data size emphasizes two aspects: performance issues, requiring faster computations of data, especially when speaking of interactive analytics; information overloading, i.e. the larger amounts of data to be presented to the user might require stronger filtering, simplification and abstraction procedures in order to provide him/her an understandable (and as much accurate as possible) concise representation of what was in the original data.

The scalability issues met in several data domains put emphasis to the recent trends in the development of **new general computational models**, especially those pushing towards the distribution of the computation over large pools of commodity computers, such as MapReduce-based solutions or the emerging Cloud Computing, just to mention two popular examples. These technologies already yielded some major success stories, yet they are deemed to be still far from mature and to suffer from drawbacks that range from data transfer bottlenecks to performance unpredictability and security (in the case of Cloud Computing), or to miss important features that traditional DBMS's provide, such as the lack of indexing which would make MapReduce-like approaches heavily based on brute force processing. Moreover, while such solutions allow efficient solutions for some large application scenarios, specific data types and specific problems might not fit satisfactorily with such general platforms and paradigms. On this direction, some examples of data-specific solutions are being proposed in literature, e.g. the Giraph graph processing system [13], also used by Facebook, based on the MapReduce implementation provided by Apache Hadoop.

■ Timeliness

In several contexts **timeliness** is a strict requirement, i.e. it is not sufficient to be able to analyze the data we have, but we need to provide answers within a tight time frame. Often that simply means we are not able to execute the full analysis process we would like to have, and therefore some mechanisms need to be developed to find good and quick approximations of the results. In particular, **anytime analytical queries** are a natural candidate, that is to say query answering and analysis processes that can continuously return an approximation of the final answer, with the property that the longer is the processing time allowed, the more accurate is the approximate answer. This kind of approach covers all those applications that involve an interactive decision making process where decisions need to be taken timely, yet with room for later refinements.

As mentioned above in this section, Big Data have often a dynamic nature, and come as a **continuous stream** of information whose size does not permit to simply store it and postpone its processing to an off-line phase. As recognized in recent literature (E.g. [3]) one of the great challenges here is to find proper mechanisms to (i) reduce the data size through compression, redundancy reduction and analysis-driven summaries computed on the fly before the storage; and (ii) to manage the whole data life cycle, including an initial selection of useful information to keep vs. less useful information to be discarded, but also data prioritization methods to filter obsolete information that can be removed from storage to make room for fresher data. An alternative approach to the data explosion issue consists in simply avoiding any pre-formatting or loading of data into a database, promoting instead the efficient execution of in-situ queries on the raw, heterogeneous data. Similarly, [17] discusses database architectures aimed to big data

exploration, arguing that storing and accessing all data is not only unsustainable with Big Data, but also unnecessary, since often each specific task needs to access only tiny portions of the database. In this direction, adaptive indexing, adaptive loading and sampling-based query processing are considered as key tools towards creating dedicated exploration systems.

■ Mastering the complexity of the Knowledge Discovery process

The notion of **BDA-as-a-service** introduced at the beginning of this paper involves the definition of analytical processes that combine available data sources and a palette of analytical tools to transform raw data into answers for the end-user. In these terms, developing BDA-as-a-service requires to **restructure the classical Knowledge Discovery (KDD) process** and lift it to another level of complexity.

The standard CRISP-DM (Cross Industry Standard Process for Data Mining), widely used by data miners to describe data mining processes for any analysis problem, might not fit very well the new setting. First, the (iterative) sequence of six steps in CRISP-DM – Business & Data Understanding, Data preparation, Modeling, Evaluation, Deployment – clearly reflects a data management perspective where all relevant information can be stored and cleaned before any further manipulation. This assumption might be easily violated in all those cases where the data flow is too massive to allow an exhaustive storage (filtering/compressing data on the fly to allow that would require some awareness of the analyses expected afterward) or when there are timeliness constraints. Second, mastering the complexity of the KDD process can become hardly manageable with the *flat* approach suggested by CRISP-DM. In particular, real analytical applications can involve several levels of analysis, combining the results of various processing tools to obtain complex patterns or models, to form hierarchical dependencies among the steps performed. A sample approach that might support this setting providing a more structured compositional framework for analytical processes is Mega-modeling [18], based on the recursive composition of basic modules. Each module is fed by an input data and an input patterns streams, possibly integrated with background knowledge (also in the form of data or patterns), and outputs a set of patterns and/or a flow of data. Complex functionalities can be modeled through the composition of modules that implement simpler functionalities, in a recursive way, till the level of elementary processing tools is reached.

In complex applications, the design of an analytical process is actually a multi-disciplinary effort that involves actors with different backgrounds. For instance, biological applications will clearly need the collaboration of data analysts with biologists which are not necessarily experts in analytical tools and processes – although in the future data analysis might become an important aspect of the education of scientists outside computer science, as well as of any professional that might benefit from data analysis. Similarly, Business Intelligence applications will require a mix of analysts and specialists of the specific business involved. In order to let them cooperate in the process design, and to let each actor benefit from the BDA services provided, different layers of abstraction of the analytical process should be provided for different users. This matches with the view expressed in [1]: “Businesses typically will outsource Big Data processing, or many aspects of it. Declarative specifications are required to enable technically meaningful service level agreements,[...]”. In this direction, the existing research lines on declarative programming tools, such as Workflows [23], are expected to provide a useful background for developing ad hoc solutions for BDA.

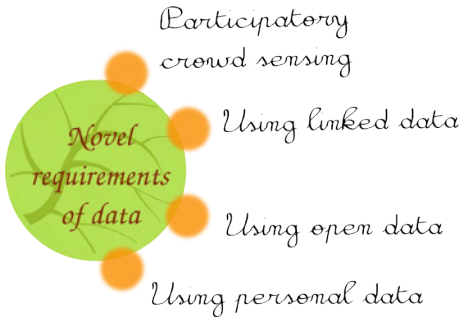
The kind of innovations requested to deal with Big Data, especially in terms of alternative architectures and computational models, call for progresses also at purely theoretical levels. In particular, the theory of computational complexity is still largely tied to classical architectures, and while it is recognized that in distributed settings some additional factors need to be considered (for

instance the obvious dependence on the number of computers involved, but also the huge impact of communications on performances), a set of suitable **complexity measures for the new algorithms** is still missing. Such set should primarily include distribution-informed complexity measures, that, where needed, can fit algorithms developed over non-classical architectures, therefore considering all the factors that reasonably have effects on the real cost of the computation.

Finally, BDA services will often involve the use of personal data, ranging from medical records to location information, activity records on social networks, web navigation and searching history, etc. All this calls for mechanism that ensure that the information flow employed in the analyses does not harm the privacy of individuals, including security aspects (tight access control to sensible information and/or data that might lead to a de-anonimization threats), anonymity guarantees, hiding sensible output patterns/models, etc.

3.2 The novel requirements of data

All the potential value that lies in Big Data can be fully realized only if a proper, efficient, fair and ethical access to data is provided to the actors involved in the service. This general objective poses not only technical questions, but also social (or techno-social) ones, since it might involve extending or reinforcing the rights of individuals and companies for matters concerning personal and public data.



■ Using personal data

Personal information constitute a significant part of Big Data, especially in the domain of social data: either passively collected through sensors (e.g. the personal location captured by a smartphone) or provided by the user in exchange of a service (e.g. credit card transactions, but also usage of location-based-services or posts on a social network), large amounts of information generated by the individual feed several different service providers, which now also play the role of data collectors.

The natural question that emerges in this setting is “who owns and should decide about the usage of such data”? An equally natural answer is given in [14, p.79]: people should have the ownership of their data, which means to have full control of how they are used and by whom, as well as the right to dispose them or to distribute them at the user's free choice. In practical terms, that requires to fully **empower the individual about his/her data**, letting him/her access and handle all of them (self-collected data, social networks, economic transactions, etc.), and decide which to share with whom – maybe donate them to public research bodies for the common benefit, or to sell them to a company.

Moving towards this scenario, where the user has a complete view and control of his/her data is not straightforward. First, as mentioned in [1], “[...] beyond record-level access control we do not understand what it means to share data, how the shared data can be linked, and how to give users fine-grained control over this sharing.” The (important) technical issues to provide access to and control of the personal data themselves are only part of the problem, since choosing to share a piece of personal information might have effects that are difficult for the user to understand. For instance, users might light-heartedly share some apparently neutral information, which however might later be linked to other data sources and allow a malicious party to associate sensible information to the user. Providing such kind of awareness to the user is a challenging task that

involves a larger view of the data ecosystem (larger than the individual's personal data only) and appropriate inference tools.

This kind of paradigm, where personal data are transparently used for analytical purposes is envisioned by several parties. For instance, in relation to USA, [21] states that “the FIPPs [United States Federal Trade Commission's Fair Information Practice Principles] should be used as a set of levers, which can be modulated to address big data by relaxing the principles of data minimization and individual control while tightening requirements for transparency, access and accuracy”. The [26] focuses more on the role played by the individual and the importance of his/her awareness: “From transparency to understanding: There is a need for new approaches that help individuals understand how and when data is being collected, how the data is being used and the implications of those actions.” and “From passive consent to engaged individuals: Organizations need to engage and empower individuals more effectively and efficiently. Rather than merely providing a binary yes-or-no consent at the initial point of collection, individuals need new ways to exercise choice and control, especially where data uses most affect them. They need a better understanding of the overall value exchange so that they can make truly informed choices.”

The last statement above from [26] also points out an aspect related to the “new deal on data” proposed in [14]: **understanding the value of personal data**, to let the user taking informed decisions. In cases where the data might be used for the social good – medical research, improvement of public transports, contrasting flu epidemics, etc. – understanding such value means to correctly evaluate the balance between public benefits and personal loss of protection. When the data is aimed to be used for commercial purposes, the value mentioned above might instead translate into a simple pricing of personal information, that the user might sell to a company for its business. This clearly might lead to ethical issues in case of misuse of such tools, for instance it is conceivable that some black markets of personal information might emerge, and therefore it calls for a proper evaluation of this kind of side effects and the realization of proper control policies.

■ Using linked data

Making sense of large data flows is the underlying common objective of any analytical usage of Big Data. However, a key requirement for that is that the data itself (i) is provided with a proper semantics, for instance in the form of meta data, and (ii) the pieces of information that have some logical connection (for instance they describe the same individual) should be linked to each other, even when they come from different data sources. Beside the technical issues involved in reconstructing semantics and links over semantically poor Big Data, already mentioned is Section 3.1, an effort is required to migrate from isolated and heterogeneous data sources to a **federation of linked data** – i.e. a system made of data objects linked to each other through relations of various kinds. This objective requires the large scale adoption of ontologies and standards for data representation and linkage by the main players of the area, i.e. data providers / collectors and data analysts/consumers. In particular, past experiences in similar efforts taught that a fully top-down approach, where such standards are enforced by some authorities and committees, are likely to fail. Therefore, we envisage a trade-off solution, where the bottom-up solutions coming from single proponents play a role in defining the guidelines to be adopted collectively.

■ Using open data

The **Open Data** movement (in consonance with the open software and other “open” movements) is expected to fuel the development of openly accessible repositories of high quality data, with clear benefits for Governments, public agencies, organizations and individuals that can use them

for their purposes. As mentioned in [19], “Despite recent advances in structured data publishing on the Web (such as RDFa and the schema.org initiative) the question arises how larger Open Data sets can be published, described in order to make them easily discoverable and facilitate the integration as well as analysis.” A problem for which [19] points to Open Data “portals” as a candidate solution that is growing in recent times. Implementing a BDA-as-a-service would clearly require to grant access to all portals able to provide information relevant for the kind of analyses expected to perform. Some efforts to promote the creation of open data repositories are in place in the EU research practice (e.g. by the Pilot on Open Research Data in Horizon 2020 [15]), yet they should be incremented significantly. One of the mechanisms that are emerging in this direction is the “**data citation**” proposed for research data repositories, i.e. the adoption of standardized referencing for datasets just as bibliographic citations for printed material, having also the effect of giving value to the creation and sharing of (open) data sets by proposing “data citations” (credited to the authors) as a valid entry in standard bibliometrics evaluation indices (H-index, etc.).

- **Participatory crowd sensing**

  Crowdsourcing tools can be used for eliciting facts, opinions and judgments from crowds of participants. Crowdsourcing systems can be adopted, for instance, to perform ad-hoc campaigns. In this context, games with a purpose can be deployed, capable of engaging people to the resolution of simple tasks, like the collection of distributed data (e.g., about urban or environment conditions) and the validation and refinement of uncertain knowledge.

3.3 Educating & promoting data analysis literacy (datacy)

While Big Data is imposing as an important element of the present and the future of our society and our lives, the society and individuals themselves apparently lack the cultural background needed to happily live with that. A poor knowledge of what Big Data actually are and how to benefit from them characterizes common citizens, who wonder what kind of personal information about them is circulating in the farms of the Big Data collectors, as well as non-specialized scholars, who refrain from dealing with Big Data because they have origins and characteristics that differ from the data sources traditionally employed in their (the scholars') specific domain.

McKinsey predicts not only that 140,000-190,000 workers with “deep analytical” experience will be needed in the US, but also that 1.5 million managers will need to become data-literate. Such a pervasive need of skills for dealing with Big Data Analytics is also evident in scientific research, as testified by the emergence, in recent years, of the so-called “4th Paradigm of Science” [16], i.e. a scientific view based on full exploitation of Big Data.

- **Data scientists**

  A significant constraint on realizing wisdom and value from big data is the shortage of talent, particularly of people with deep expertise in statistics and machine learning, and the managers and analysts who know how to operate companies and institutions by using insights from big data. Skills must be developed on how to exploit data and their analysis to develop successful business initiatives. The skill gap in data science is also a barrier to the exploitation of big data for social good: support policy making, novel ways of producing high-quality and high-precision statistical information, to empower citizens with self-awareness tools, and by promoting ethical uses of big data.

Researches in this field have identified three key types of talent required to capture value from big data:

- *deep analytical talent* – people with technical skills in statistics and machine learning, for example, who are capable of analyzing large volumes of data to derive business insights;
- *data-savvy managers and analysts* who have the skills to be effective consumers of big data insights – i.e., capable of posing the right questions for analysis, interpreting and challenging the results, and making appropriate decisions;
- supporting *technology personnel* who develop, implement, and maintain the hardware and software tools such as databases and analytic programs needed to make use of big data.

The shortage of professionals in data analysis (first item of the list above) is already well recognized as a critical point for business and research bodies. The second category of users – knowledgeable enough on what Big Data can say and on the needed analytical processes to profit the most from them, without being specialized data analysts – is now emerging as the missing link between theory and practice, i.e. between analytical methods and a useful exploitation of the data in real applications.

The education of data scientists is a challenging task, requiring to put together solid technical competences, ability to narrate the stories that data tell after analysis and modeling (e.g. using both visual and multi-media storytelling), and preparation towards the various ethical and legal issues connected to Big Data Analytics and the management of data pertaining to people. Also, BDA is today a fluid research arena, with novel projects and use cases everyday, and therefore it is challenging to enucleate, from this dynamic state-of-the-art, the syllabus of fundamental concepts and techniques that should constitute the data scientist's background.

- **Datacy**

Given the pervasiveness of Big Data in most of the disciplines of human knowledge and research, elements of data science should be provided to students of all levels of education, from high-schools to university curricula. At the lowest levels, that would help to improve the general awareness of importance and value of the data (including scientific data but also personal information, etc.) for science and business, as well as a view of what data science can do with them. We can call this kind of education **datacy** in assonance with literacy, to emphasize the fact that it should include the basic knowledge needed also to common citizens to comfortably live in a big data society. At the highest levels of education, that can provide future scholars and business professionals a larger perspective on methodologies and areas of exploration that might be exploited in their job. For instance, students in biology might benefit from a curricula that combines traditional methods in epidemics study and monitoring with data-driven approaches, such as Google's Flu Trends [10] and Flu Near You (an initiative run by Boston Children's Hospital).

3.4 Achieving Big Data Analytics-as-a-service

Letting BDA-as-a-service have a chance to flourish in Europe requires not only technical advancements to make it capable of providing the needed functionalities, but also several organizational actions aimed to promote development along some key directions. Such actions should be performed at the European level, and the support and leadership of the European Community would clearly be fundamental. A first list of directions is the following.

- An important goal of BDA-as-a-service should be to grant access to Big Data Analytics for small companies that cannot afford to develop a in-house platform. That would mitigate the present great unbalance between the dominating large companies and the smaller players of the market, promoting a more even distribution of the expected benefits of BDA. The key mechanism for achieving that is funding and **supporting the creation of Big Data Analytics centers** accessible to researchers, public administrations, medium and small companies. In the BDA-as-a-service spirit, such centers should make it possible and affordable to have access to large (federated) data repositories and appropriate analytics services, as well as receive the assistance of BDA specialists to exploit them efficiently. In Europe there are already some examples of federated structures mainly aimed at gathering and integrating the data sources for a given objective (E.g. the French society Data Publica [7]), yet comprehensive infrastructures that support the whole BDA process from data acquisition to the data analysis are still missing.
- Since BDA-as-a-service must rely on a federation of data repositories (or "portals" as in [19]) to be harvested for locating the information relevant to each analysis, it is fundamental to develop capabilities for cooperating, searching and exchanging data across repositories. In this direction, the EU should incentivize the adoption of a **layered framework to increase interoperability** across the single data repositories. This line of action should be a natural continuation of the efforts already put by the EU in initiatives such as Research Data Alliance [22], OpenAIRE [2], EUDAT [9]
- Thanks to its traditions of norms for the protection of individual data and its efforts to keep them aligned to the digital revolution, Europe might and should reach a **leadership in the development of privacy-enabling solutions**. The General Data Protection Regulation (GDPR), which unifies data protection within the European Union and is planned to be adopted from late 2014, constitutes a first important step towards norms that clarify the boundaries of personal data usability and that define rights for the individuals about the control of their data. GDPR includes elements such as the adoption of Privacy by Design (and by Default) principles in business processes, the right to erasure (a variant of the "right to be forgotten" or "right of oblivion") of personal information related to search engines, and the extension of the protection norms to cover all European citizens whoever is the data controller – i.e., multinational or non-EU based corporations treating data of EU citizens will be affected as well [12]. This competitive advantage over the other most active countries on Big Data (USA and Asian countries) should be maintained and leveraged for the growth of Big Data Analytics in Europe.

4 Deployment scenarios

A most effective way for promoting and helping the development of Big Data Analytics is to create successful, large-scale showcases. In particular, high-impact application domains should be given priority, to maximize visibility to the public and to prove the value towards a large pool of stakeholders, e.g. policy makers. The following list provides just a short selection of examples, which might be easily extended including several business or social good-related applications.

- **Smart cities and communities**. Being one of the hot topics of the H2020 research agenda, it is natural to include it as a priority showcase. Modern cities are the perfect example of environment that is densely traversed by (mostly user-generated) large data flows: traffic monitoring systems, environmental sensors, GPS-enabled individual mobility traces, city-related information posted on social networks, etc. On the other hand, cities are the stage of a large-scale collective sharing of resources that needs to be optimized and continuously monitored and promptly adjusted when needed: urban planning, public transportation, reduction of energy consumption, ecological sustainability, safety and management of mass events are just the front line of topics that can benefit from the awareness that Big Data might in principle provide to the city stakeholders. All

such keywords point to applications for the social good, that indirectly translates into benefits for the individual in the form of improved public transports, a safer and healthy living environment, etc. These can be also extended with other applications aimed directly to the individual, for instance as services that improve his/her awareness (the global traffic status and forecasts, nearby events that might fit my recreational interests, etc.) or that answer to specific requests (urban journey planners, finding the most convenient places to go for performing a given activity avoiding crowds, etc.). Finally, there is also room for a market of business applications, such as real-time targeted marketing (proposing the right offer to the right person at the right moment, maybe also in the right place), paid services to the individual, Big Data-based market researches, etc. This application context emphasizes two important aspects of BDA: the need (and room) for creativeness to exploit and combine the several data sources in novel ways; the need to give awareness and control of the personal data to the users that generate them, in order to sustain a transparent crowd-sourced data ecosystem that feeds the applications.

- **BDA for developing countries.** The impact of BDA-driven applications can be greater in cases where they help to build new infrastructures or new services to the population, or to make poor existing ones perform a leap forward to a different quality level. Helping the growth of developing countries might represent an exemplary case: on one hand, BDA can help making decisions right at the planning stage, or at an early development that can be still deeply influenced by a better understanding of population needs (through inference of mobility demand, detection of requests for – or negative feedbacks about – specific services, etc.) and country/urban status (traffic congestion, distribution of energy consumption, over/under-usage of resources, etc.). On the other hand, Big Data is in some cases the only source of information about the country/city available, lacking some forms of official data or surveys. A sample effort in this direction is the Data for Development (D4D) initiative [6], where mobile phone data of a developing country has been shared by a telecom operator with researchers expressively to develop analytical tools and applications for the social good of the country. The outcome was a long list of solutions that provided several insights about transportation, social studies (especially about migration flows and social divisions), disease containment strategies, and so on. Another example is provided by the United Nations Global Pulse project, that envisages a future in which big data is harnessed safely and responsibly as a public good. Its mission is to accelerate discovery, development and scaled adoption of big data innovation for sustainable development and humanitarian action [25].

- **Global market of jobs.** Searching for a job has traditionally been a matter of finding a proper match between the skills and requirements (or expectations) of an individual and the features of job offers available in his/her neighborhood. With the growth of Europe as a unified entity and the increasing ease of moving across countries, the search is expanding its boundaries throughout the continent. The job vs. worker matching is typically reduced to find a pairing between under-specified calls for works and a curriculum vitae that roughly approximates some of the individual skills expected to be (by the individual him/her-self) relevant for the work offers. The various levels of approximation and missing information on both ends of the pairing can easily lead to poor results. Also, job seekers and job proponents usually have an extremely incomplete vision of each other, i.e. a globally shared job marketplace does not exist, yet. The basic objectives of BDA, on the other hand, include the ability to collect and link together large pieces of information related to the same individual or the same topic, forming a coherent holistic view of it/her/him. That can basically lay the basis for developing smart profiling tools able to objectively reconstruct skills, experiences and attitudes of job seekers, in a way that goes much beyond standard curricula. Also, such mechanisms might even allow to make the process proactive, letting the job proponents take the initiative in finding good candidates. Finally, services of this kind might also look for conditional pairings, i.e. also include job offers that require some missing skills, which yet might be integrated by the user in his/her curriculum through a moderate effort in education – e.g. having a short specialization course.

- **Quantitative assessment of results of European projects and activities**. Big Data give the chance to measure and understand several aspects of individual life, urban and country development, economic and health status, as well as their changes in time. As such, they have the potential to provide a framework of objective indicators of the impact of initiatives that are expect to have (usually beneficial) effects on society at large. For instance, the success of a project that promises to make cities smarter and greener, might be measured not only through an evaluation of academic results and subjective opinions provided by experts of the field, but also by comparing various indicators of city-smartness and city-greenness – e.g. percentage of kilometers performed by citizens on private cars vs. public transport, estimates of CO_2 emissions, average time spent waiting for bus connections, energy consumption levels. Clearly, applications that provide a complete range of indicators of this kind might result extremely helpful for policy makers, for instance to provide quantitative assessment of the forthcoming wave of results of H2020 activities and projects.
- **Big Data-aware Official Statistics**. According to [24], Official Statistics "provide an indispensable element in the information system of a democratic society, serving the government, the economy and the public with data about the economic, demographic, social and environmental situation". Big Data means to bring new types of data and new data sources with particular characteristics into a well defined information production process. Beside the critical issues emerging when trying to integrate these two elements (Big Data vs. traditional Official Statistics processes, some of which were discussed in previous sections), Big Data show the potential to improve Official Statistics processes in various important ways: first, the kind of information now produced by Official Statistics might be extracted in a cheaper and more timely manner, for instance substituting (part of) current surveys with alternative and streaming data sources; second, small-scale phenomena that traditional statistical instruments fail to properly capture might be better described by Big Data, for instance by using mobile phone activity to infer the real portion of population that actively uses a territory; finally, Big Data enable the measurement of phenomena previously inexistent (for instance, usage of social networks or online services for the citizens) or near to impossible to capture (for instance, the mood and reactions of the population caused by unexpected events or important news, now possible to some extent through the analysis of social media and human mobility data sources).

As well known by data analysis practitioners, dealing with large real applications is also an extremely proficient way to understand which analytical problems are relevant and important, and therefore to update the priorities in – or redefine – the technical research agenda accordingly.

5 Economic and business perspective

Beside providing potential benefits for science and society, Big Data might become a big asset European Union's Economy. Big Data Analytics can be the means for optimizing at least the following major constraints in economic activity:

- **Capture and predict customer demand**. Consumer demand is a fundamental constraint to the architecture of the free market system. Obviously, if we remove this constraint then production is no longer driven by the demand of paying customers and we no longer have a free market economy. Big Data Technology can help us to capture and predict this demand either for a single customer (segment size = 1) or for the entire customer group (segment size =all), with high accuracy and in "near real time", with respect to the production process.
- **Optimize industrial production**. In today's economy, and not only of the European Union, human labor is required to some degree in the production of nearly every product. Big Data Technology can help to (semi-)automatize less complex and routine jobs, as well as provide effective support

for more complex ones, such as optimizing supply chain management, monitoring product quality or determining optimal sales approaches, among many other areas.

- **Optimize consumption of energy and of other production resources**. Europe's production is limited by resources, such as energy, land, water etc. In addition, production causes external effects on the environment, such as toxic pollution, use of public resources, climate change effects etc. Already Big Data Technology helped in optimizing logistics, analyzing weather data, improving public transportation, better matching electrical energy production with its consume, etc.
- **Optimize technology development**. Obviously, the sophistication of the technology, machines and products limits production. As technology advances, it is no longer a tool that requires workers to tune it but can act more and more autonomously. For example, relational data base technology, which is at the core of each IT system in industry, required till the early 2000s highly paid IT-experts. These days, relational databases are mainly self-regulated. Algorithms take care of optimizing the complex technology for data management. Hence, a long term goal is a (nearly) self-regulating production of technology.

Optimizing these major constraints with what we now call Big Data Technology is not new. Since the 90ies we observe this technology creating so called superstars in winner-takes-all markets, such as Web search (Google), logistics (Amazon) or social networks (Facebook). Three fundamental technical building blocks are still the main impact factors for the above constraints in organization's economic activity:

- **Continuous Digitalization Streams for Europe's Industry**. Cheap data storage on hard discs is now available to many industries, giving them an incentive to store huge amounts of rather raw data in a variety of data sources – with the hope to utilize this data later for optimizing the above mentioned four constraints in their economic activity. Since several years now an inexpensive and robust file system, the Hadoop Distributed File System and extensions such as Parquet.io or HBASE, is available under an industry friendly open source license. These file systems permit also small and medium sized companies storing valuable information from the enterprise data warehouse and other systems in a single and robust storage system. Data variety, data distributed on several sources and unknown data quality are some of the major challenges in this context. Typically digitized data is poorly linked to each other, even within the same organization. All problems related to data integration and linking discussed in Secs. 4.1 and 4.2 directly apply here.
- **Interactive Pattern Recognition – Not only for Data Scientists**. In the ideal world, business owners could directly recognize patterns in core business processes, such as of customer's behavior, production and development, resource consumption and sales. Contrary, in today's world recognizing patterns in data is a cumbersome process of collecting, wrangling, cleansing, profiling and modeling data that requires to accompany business officers with specifically trained data scientists having both a business mindset and a deep technological understanding of data. The great challenge is therefore to provide Big Data Technology for enabling and simplifying interactive pattern recognition for 'everybody', in particular for business owners in non-IT and small and medium sized companies.
- **Ideation and Value Creation from Data for Entrepreneurs**. The ideas of the previous two points is to make available massive bodies of data to almost any situation and permit any business owner to recognize and to rank shortcomings in business processes, to spot potential threads and win-win situations. Ideally, "every" European citizen could establish from these patterns new business ideas. Going further along this direction, we envision a world where Big Data Technology helps European entrepreneurs to recombine ideas and previous innovations into new services and products, as well as to go through all these potential idea combinations and select the truly valuable ones.

6 Conclusion and recommendations

In this document we discussed several points that, according to the authors, will have a major relevance for the development of BDA in Europe in next years. We conclude with a few, high-level recommendations that summarize most of the key points. An important premise is that in the last 10 years European research invested a lot on Database and Data Mining technology, and managed to develop a strong base of expertise and innovation on these topics, therefore we believe that future actions should capitalize and advance on that base.

- Recommendation 1: EU should spawn federations of key public and private actors, in **challenging multidisciplinary** domains to provide a critical mass for starting up Federated Trusted Open Analytical Platforms for Knowledge Acceleration and creating incentives for further actors to join.
- Recommendation 2: In EU should **support the creation of Big Data Analytics centers** accessible to researchers, public administrations, medium and small companies.
- Recommendation 3: Funding and supporting the development of the **technologies** needed to empower citizens, public institutions and businesses based on the values of the Federated Trusted Open Analytical Platforms for Knowledge Acceleration.
- Recommendation 4: Promoting the development of a **normative framework** for the above mentioned empowering of citizens, public institutions and businesses along four dimensions: privacy-preservation, trust management, individual empowerment and public good.
- Recommendation 5: Promoting **education** of novel data scientists and ***datacy***.
- Recommendation 6: Promoting incentives for providing **data access** to researchers, business actors and public administrations. Examples include assigning rewards to and/or facilitating virtuous business actors that share and maintain open data portals; giving value to “data citation” in research, i.e. recognizing the citations to a data collection (and therefore to the people that collect, curate and document it) as a valid bibliometrics indicator; and enforcing regulations. This line of action should be a natural continuation of the efforts already put by the EU in initiatives such as Research Data Alliance [22], OpenAIRE [2], EUDAT [9].

7 Workshop participants, organization and methodology

A group of twenty researchers from the areas of core database technology and data analytics met in Pisa, Italy in May 2014 to discuss the state of research in these areas in the context of Big Data, its impact on practice or education, and important new directions. The attendees represented a broad cross-section of interests, affiliations, seniority, and geography.

Before the workshop, each participant submitted a short paper summarizing his/her vision of Big Data Analytics in the next 10 years, the main challenges to be addressed according to this vision, and the research directions that should to be undertaken in the next future to properly confront such challenges. Such papers have been shared among the participants, in order to set the stage for a more effective discussion and comparison of different perspectives.

During the workshop two subgroups were formed: one mainly formed by specialists on database technologies, which focused on the **data management issues** introduced by Big Data; and one mainly formed by researchers on data analytics, which gave emphasis to the new **analytical opportunities** opened by Big Data, together with the research issues they give rise. The two groups alternated separate meetings/brainstorming sessions to plenary meetings to share results and keep the overall efforts focused. The workshop terminated with the collaborative preparation of a declaration of aims and an index of context that were later expanded remotely to obtain the present paper.

Workshop organizers:

- Fosca Giannotti, ISTI-CNR
- Mirco Nanni, ISTI-CNR **(document editor)**
- Andreas Rauber, TU Wien
- Costantino Thanos, ISTI-CNR

List of participants and contributors:

Data Management Group
Anastasia Ailamaki, EPFL
Peter Baumann, Jacobs University
Martin Kersten, CWI
Alexander Löser, Beuth Hochschule
Andreas Rauber, TU Wien
Nicolas Spyratos, Paris-South – Orsay
Costantino Thanos, ISTI-CNR

Data Analytics Group
Emanuele Baldacci, ISTAT
Francesco Bonchi, Yahoo! Labs
Fosca Giannotti, ISTI-CNR
Stan Matwin, Dalhousie University
Dunja Mladenic, J. Stefan Institute
Mirco Nanni, ISTI-CNR
Amedeo Napoli, LORIA
Dino Pedreschi, University of Pisa
Michalis Vazirgiannis, LIX Ecole Polytechnique

Contact person: Mirco Nanni, ISTI-CNR, mirco.nanni@isti.cnr.it, Phone: +39-050-6212843.

7.1 Acknowledgements

This work has been proposed and pursued as an activity of the SoBigData Lab in answer to the ERCIM call for the creation of expert groups on “Big Data Analytics”. ERCIM supported all the process, funding the workshop held in Tirrenia (Pisa, Italy) and providing technical infrastructures for document sharing and collaboration within the working group.

8 References

[1] Agrawal et al. Challenges and Opportunities with Big Data 2011-1 (2011). Cyber Center Technical Reports. Paper 1. http://docs.lib.purdue.edu/cctech/1

[2] The OpenAIRE project. https://www.openaire.eu/

[3] Big Data: A Survey. Mobile Networks and Applications. April 2014, Volume 19, Issue 2, pp 171-209. Springer, http://link.springer.com/article/10.1007%2Fs11036-013-0489-0.

[4] Framing a European Partnership for a Big Data Value Ecosystem. EU Projects Big and Nessi report. February 2014.

[5] Big Data Privacy Workshop: Advancing the State of the Art in Technology and Practice. The White House Office of Science & Technology Policy and MIT. March 3, 2014.

[6] http://www.d4d.orange.com/en

[7] http://www.data-publica.com/

[8] Data, data everywhere. The Economist, Special Report on Big Data, February 2010.

[9] EUDAT. http://www.eudat.eu/

[10] http://www.google.org/flutrends/

[11] Federal Trade Commission. Big Data: A Tool for Inclusion or Exclusion?. September 2014. http://www.ftc.gov/news-events/events-calendar/2014/09/big-data-tool-inclusion-or-exclusion

[12] European Commission – Justice – Data Protection. Reform of data protection legislation. June 2014. http://ec.europa.eu/justice/data-protection/

[13] Apache Giraph. http://giraph.apache.org/

[14] World Economic Forum: The Global Information Technology Report 2008–2009. Mobility in a Networked World. www.weforum.org/pdf/gitr/2009/gitr09fullreport.pdf

[15] Guidelines on Data Management in Horizon 2020. December 2013. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[16] A. J. Hey, S. Tansley, K. M. Tolle, et al. The fourth paradigm: data-intensive scientific discovery. Microsoft Research Redmond, WA, 2009.

[17] Idreos S. Big Data Exploration. In: Big Data Computing. Taylor and Francis; 2013.

[18] Ceri, Della Valle, Pedreschi, Trasarti. Mega-modeling for Big Data Analytics. Lecture Notes in Computer Science, v. 7532, 2012, pp 1-15. http://link.springer.com/chapter/10.1007/978-3-642-34002-4_1 .

[19] NEM Networked and electronic media. Big and Open data. Position Paper. December 2013. http://nem-initiative.org/wp-content/uploads/2013/11/NEM-PP-016.pdf

[20] National Science Foundation, Solicitation 12-499: Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA), 2012

[21] Tene, Omer and Polonetsky, Jules, Big Data for All: Privacy and User Control in the Age of Analytics. Northwestern J. of Tech. and Intellectual Property, 239 (2013). http://ssrn.com/abstract=2149364

[22] Research Data Alliance Europe . Report on the RDA-MPG Science Workshop on Data . April 2014. https://europe.rd-alliance.org/Repository/document/Publications%20and%20Reports/RDA-Europe-Science-Workshop-Report_final_April2014.pdf

[23] I. Taylor, E. Deelman, D. Gannon, M. Shields (eds) (2007) Workflows for e-Science. Springer, New York, Secaucus, NJ, USA.

[24] UNECE, “What does “Big Data” mean for official statistics?” Conf. of European Statisticians (2013) .

[25] United Nations Global Pulse. http://www.unglobalpulse.org/

[26] World Economic Forum. Unlocking the Value of Personal Data: From Collection to Usage. 2013.

[27] The White House. Big Data: Seizing Opportunities, Preserving Values. May 2014. http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf