International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-018-0126-x

REGULAR PAPER

Data science at SoBigData: the European research infrastructure for social mining and big data analytics

Valerio Grossi[1] · Beatrice Rapisarda[2] · Fosca Giannotti[1] · Dino Pedreschi[3]

Received: 20 March 2017 / Accepted: 27 April 2018

Abstract
Most people have become "big data" producers in their daily life. Our desires, opinions, sentiments, social links as well as our mobile phone calls and GPS track leave traces of our behaviours. To transform these data into knowledge, value is a complex task of data science. This paper shows how the SoBigData Research Infrastructure supports data science towards the new frontiers of big data exploitation. Our research infrastructure serves a large community of social sensing and social mining researchers and it reduces the gap between existing research centres present at European level. SoBigData integrates resources and creates an infrastructure where sharing data and methods among text miners, visual analytics researchers, socio-economic scientists, network scientists, political scientists, humanities researchers can indeed occur. The main concepts related to SoBigData Research Infrastructure are presented. These concepts support virtual and transnational (on-site) access to the resources. Creating and supporting research communities are considered to be of vital importance for the success of our research infrastructure, as well as contributing to train the new generation of data scientists. Furthermore, this paper introduces the concept of exploratory and shows their role in the promotion of the use of our research infrastructure. The exploratories presented in this paper represent also a set of real applications in the context of social mining. Finally, a special attention is given to the legal and ethical aspects. Everything in SoBigData is supervised by an ethical and legal framework.

Keywords Research infrastructure · Social mining · Social sensing · Data science

1 Introduction

Many aspects of our daily activities generate facts that can be stored and analysed. Our social life leaves traces in the network of our phone calls or email, in the friendship links of our social networks; our shopping patterns and lifestyles are saved in the records of purchases; our movements are stored in the records of our mobile phone and GPS tracks. Multiple aspects of our social life are "big data" proxies. In this complex and evolving scenario, the role of data science is to transform data into knowledge and value. Furthermore, it is important to highlight that a significant barrier to realize knowledge 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 should 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, e.g. supporting 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.

For this reason, SoBigData Research Infrastructure (RI) serves a large community of social sensing and social mining

✉ Valerio Grossi
vgrossi@di.unipi.it

Beatrice Rapisarda
beatrice.rapisarda@iit.cnr.it

Fosca Giannotti
fosca.giannotti@isti.cnr.it

Dino Pedreschi
pedre@di.unipi.it

[1] ISTI-CNR Pisa, via Moruzzi 1, 56124 Pisa, Italy

[2] IIT-CNR Pisa, via Moruzzi 1, 56124 Pisa, Italy

[3] Dipartimento di Informatica, Universitá di Pisa, Largo B. Pontecorvo, 3, 56127 Pisa, Italy

Published online: 15 May 2018

researchers. Social sensing includes techniques for collecting the digital footprints generated by humans when interacting with the techno-social ecosystem and for making such data available for mining/analysis at properly high abstraction levels. Social mining provides the analytical methods (and associated algorithms) needed to understand human behaviour by means of automated discovery of the patterns, rules and profiles from the massive datasets of human activity. Currently, this expertise and tools are fragmented, due to the difficulties in sharing experimental data and methods. In practice, the big data research revolution is in its infancy, but there is a gap between the opportunities and the status quo. The SoBigData RI aims to integrate these resources and create a framework where sharing data, methods, ideas and expertise among scientists and researchers can occur.

The SoBigData community includes scientific, industrial, and other stakeholders. In particular, our stakeholder analysis shows that there are data analysts and researchers (35.6%), followed by companies (33.3%) and policy and law makers (20%) [21]. For the first time in social mining, researchers and industrial users have the opportunity to access a rich social data ecosystem in a unified manner, and to conduct large-scale, multi-disciplinary experiments on extracting social knowledge. Furthermore, there is a large potential for developing new affordable, big social data analytics products and services, since many companies in diverse areas are analysing and comparing big social data, often in a labour intensive and expensive manner. For this reason, the need of computational and storage resources together with the capacity to preserve the results of analysis performed is key points for increasing the services provided by our RI and guarantees a continuous growth of the social mining community not only in terms of number of users, but also for the quality and the speed of our research and analysis. Several SoBigData services are aimed at constructing the community composed by excellence centres, other academic and industrial users, and trainee data scientists.

Finally, interoperability[1] is among the most critical issues to be faced when building a new system as a "collection" of independently developed constituents (systems on their own) that should cooperate and rely on each other to contribute to implement the system tasks.

Paper organization: Section 2 introduces the main pillars whereon SoBigData RI is based and the services it provides. Section 3 shows how SoBigData RI supports data science. Finally, Sect. 4 draws some conclusions and provides a bridge towards the future.

2 SoBigData: resources and access opportunities

SoBigData serves a cross-disciplinary community of data scientists studying all the aspects of societal complexity from a data- and model-driven perspective.

Figure 1 shows the main concepts related to SoBigData RI.

On the one hand, our RI provides a set of services to the research community by means of different types of access to the resources and laboratories. The "impact on society" of these services is based on a set of key performance indicators that show for example how many people use our services and take advantage of our Virtual and Transnational Access[2] or how many people attend our dissemination or training events.

On the other hand, all the services provided by RI are based on a solid range of social mining techniques and employ different kinds of dataset including open data, virtual collection by means of web crawlers and scrapers and restricted dataset that cannot be freely downloaded but the access is on-site and supervised by RI experts.

Finally, since we are often in the presence of personal and sensitive data describing human activities, every activity under SoBigData is supervised by an ethical and legal framework.

The following subsections describe the concepts introduced in Fig. 1 starting from the services provided by the RI (Sect. 2.2) and showing its main components (sections from 2.3 to 2.5).

2.1 SoBigData management bodies

Before introducing the key components and services of the RI, we introduce a short description of the main management bodies inside SoBigData. Figure 2 shows all the bodies inside SoBigData. Some of them are for administrative issues related to H2020 programme, and then specifically requested by European Community (EC), while others are for managing specific aspects and services provided by RI.

Recalling Fig. 1, it is important to understand that the SoBigData management is a complex task, since it involves both tasks performed for developing the RI (for managing the project as defined by the description of work) and actions for managing users and services towards community. In this context, we outlines only the most important ones in order to enable a clearer view of the overall infrastructure clarifying also how some concepts that will be introduced in the next sections have been defined.

[1] The IEEE Glossary defines interoperability as "the ability of two or more systems or components to exchange information and to use the information that has been exchanged" [1].

[2] The formal definition of Virtual and Transnational access (and their key performance indicators) is defined by the European Community—Infraia-1-2014-2015 call (https://goo.gl/E6Cyze).

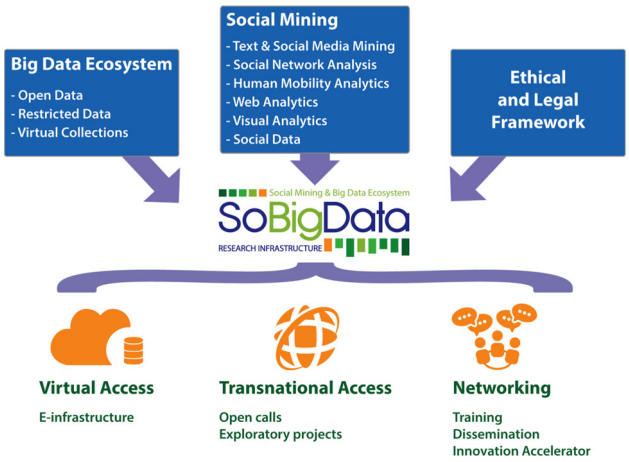

Fig. 1 SoBigData resources and access opportunities: on the higher side, we have our engines related to Social Mining and Big Data management. On the lower one, we have the SoBigData services for promoting data science by means of virtual and transnational access and networking events. All the activities inside the RI are supervised by an ethical and legal framework

Fig. 2 SoBigData management boards

Project steering board: this board is the executive decision-making body. Each item defined in Fig. 1 specifies its technical inputs, progress and effort expenditure within the respective work package and this board balances these inputs, along practical lines, within the overall technical direction agreed in the description of work. Ethics and gender issues are also monitored by this body.

All the thematic areas defined (and eventually updated) inside the project are managed by this board. As shown in Sect. 3, the concept of exploratory project exploits and binds these areas. SoBigData has six main thematic areas defined at proposal time and inspired both by H2020 call and by the competencies inside the project consortium.[3] The thematic areas include the following topics: *text and social media mining*, *social network analysis*, *human mobility analytics*, *web analytics*, *visual analytics* and *social data*.

[3] The description of the project consortium is available at the following link: http://project.sobigdata.eu/consortium.

Ethics board: this body guarantees the good governance of data; the research integrity and academic ethics; the compliance with ethical and legal framework and the consultation with external experts when and where required. This board includes professors for legal informatics and IT law and acts as the legal counsel to the project steering board. This body leads and defines all the specific actions inside the legal and ethical framework shown in Sect. 2.5.

Transnational access management board: required by EC, it is responsible for the call definition, publication and advertising, for the management and setting up the international user selection panel. More specifically, each year the project manages two calls where the SoBigData RI invites researchers and professionals to apply to participate in short-term scientific missions to carry forwards their own big data projects. These applications are opened every six months. The applications come from individuals with a scientific interest, professionals, startups and innovators that benefit from training in data science and social media analytics.[4] See Sect. 2.2.1 for detailed description of TA services.

Advisory board: this body comprises international experts standing from a range of academic and non-academic organizations. Members are senior researchers, business executives, educators and decision makers from both technical and non-technical backgrounds. It is required by EC.

This board reviews "Virtual Access Service Provision and Operation (see Sect. 2.2.1 for Virtual Access definition)" reports and produces the assessment documents for EC, helping to identify problems and possible remedies. Furthermore, it helps the project steering board in technical advances of the project in a wider scientific and commercial contexts. Since the members of this board are international experts outside the SoBigData RI, they are invited to discuss also critical decision inside RI and project activities.[5]

2.2 Research infrastructure: services and access

Our RI proposes an operational Hybrid Infrastructure that: (i) supports the creation and management of Virtual Research Environments[6] (VREs) by dynamically acquiring the resources (data, tools, services, computing) from a

[4] The access of users not working in a EU or associated country is limited to 20% of the total amount of units of access provided under the grant.

[5] The list of international experts inside the SoBigData advisory board is available at the following link: http://project.sobigdata.eu/management-bodies/project-advisory-board.

[6] The VREs [5] are web-based, community-oriented, comprehensive, flexible, and secure working environments. They are conceived and tailored to satisfy the needs of a designated community. Generally, they offer: (i) a rich array of services for data discovery and access, (ii) a data analytics platform, (iii) collaboration-oriented facilities enabling scientists.

resource space, (ii) is designed to integrate resources from other e-Infrastructures and (commercial) vendors/providers by using a "system of systems" strategy, (iii) is offering unified access to the integrated resources by abstracting from the underlying e-Infrastructures, (iv) is designed to maximize resources exploitation and minimize operational costs. All these aspects are exploited for supporting community networking (training and innovation activities), data science, and for granting the access to all the SoBigData resources.

In this perspective, SoBigData provides an e-Infrastructure[7] that is "open" and "aggregative" by design, i.e. it is conceived to aggregate into a unifying resource space resources coming from many and heterogeneous providers. SoBigData RI offers a common ground for hosting the domain-specific resources and dynamically building and operating VREs offering specific and web-based working environments to target communities.

In order to be able to serve the needs of the social mining community, it is important to invest effort in adapting and extending the resources currently owned by the SoBigData community thus to let them benefit from the e-Infrastructure capacity. Existing resources are integrated into a unified space. Depending from the "level of integration" that is achieved for each resource it will be possible to support and guarantee a diverse level of management ranging from a simple discovery to their re-purposing to better serve the needs arising in a specific VRE.

The following subsections describe how the access to the resources is grated and what are the network activities supported.

2.2.1 Virtual and TransNational access

The access to the VREs is guaranteed through two kind of accesses: Virtual Access (VA) and Transnational Access (TA).

The objective of the VA is to offer online services for big data and social mining research. To this aim a portal and e-infrastructure are built up integrating and upgrading existing infrastructure in several research contexts. In this framework, the user can access all the services provided by the infrastructure using the SoBigData e-infra which is equipped with several social tools for creating and supporting virtual multidisciplinary community of researchers. Furthermore, VA is supported by the *SoBigData Lab* VRE, an environment where the user can perform experiments using distributed computational and cloud facilities, without the need of installing or downloading anything.

On the other side, TA requires an on-site access provided by the seven research centres (installations) of SoBigData in Europe. TA consists of short-term visits of researchers (or teams) at one of the installations that provides big data computing platforms, big social data resources, and cutting-edge computational methods. The aim is to give to researchers the opportunity to interact with the local experts and discuss research questions, to run experiments on non-public datasets and methods and to present results at workshops/seminars.

Both VA and TA are services for disseminating, promoting, and implementing data science as driving force for growth not only for research community but also for the impact on society and innovation. Section 3 introduces the principal working environments defined through VREs supporting VA and TA. These working environments (called exploratories in their high-level abstraction) are also important to define strategies for implementing our dissemination and community networking strategies.

2.2.2 Networking, training and dissemination

SoBigData promotes networking, dissemination and innovation actions by means of workshops, summer schools, datathons, training courses and knowledge transfer by means of industrial partnerships. These networking actions aim at bringing together not only scientists from diverse research communities but also industries and other stakeholders, e.g. policy makers, government bodies, non-profit organizations, funders.

In particular, the training events (and training resources) are aimed at creating a new generation of multi-disciplinary social data scientists. Innovation activities, including partnerships with industry, are key factors for measuring impact and sustainability of our RI. For these reasons, SoBigData supports directly a master in big data[8] and a Ph.D. course[9] at University of Pisa and Scuola Normale Superiore. Training materials resulting from the RI activities are gathered and used by the academic partners as part of their course materials at postgraduate level, as well as, used by all partners to provide courses to companies and other non-scientific stakeholders.

In its first two years of activities, the RI has organized two workshops and one conference, one summer school, two datathons and several training events also for high-school students.[10] Furthermore, the SoBigData RI has promoted two important initiatives towards industry implementing the knowledge transfer from research to companies (as requested

[7] The SoBigData e-Infra is powered by D4Science [6].

[8] The master in big data of the University of Pisa is an annual course to become data scientists (http://masterbigdata.it/en).

[9] The Ph.D. in Data Science is aimed at educating the new generation of researchers that combine their disciplinary competences with those of a data scientist (http://phd.sns.it/it/data-science/).

[10] The updated list and the description of all dissemination and training events inside SoBigData is available at the following link: http://www.sobigdata.eu/events/.

by EC) and enabling an actual innovation acceleration for big data exploitation.[11]

2.3 BigData ecosystem

This aspect is related to the definition of a distributed data ecosystem for procurement, access and curation of big data, to underpin social data mining research within an ethic-sensitive context. Our RI defines policies that guide partners and users in the collection, description, preservation and sharing of their data sets.

We recall that research on social mining relies on massive data sets of digital traces of human activities. Many big data sets are already available at the SoBigData RI including network graphs from mobile phone call data, networks crawled from many online social networks, including Facebook and Flickr, transaction micro-data from diverse retailers, query logs both from search engines and e-commerce, society-wide mobile phone call data records, GPS tracks from personal navigation devices, survey data about customer satisfaction or market research, large Web archives, billions of tweets, and data from location-aware social networks.

SoBigData makes such data available for collaborative research by adopting various strategies, ranging from sharing the open data sets with the scientific community at large, to share the data with disclosure restriction within the consortium, also on a bilateral basis, or allowing data access within secure environments at each local installation. The data access offered in SoBigData concerns both existing and newly collected datasets. The access trough VA is granted for all those datasets whose policies allow open diffusion; conversely, for all the data sets whose access is restricted due to licensing restrictions, access is provided only through TA access.

Currently (April 2018) SoBigData catalogue contains more than 150 resources including 71 methods and 74 datasets.[12]

2.4 From data to knowledge: social mining and sensing

One of the most pressing and fascinating challenges is understanding the complexity of our interconnected society. If in the previous section we highlighted the importance of data, in this one we outline the main methods available in SoBigData in order to transform data into knowledge.

As suggested by the full name of our RI—SoBigData: SOcial mining and Big Data ecosystem—the role of methods related to social mining (and sensing) is crucial for the definition of our services.

Social sensing: these kinds of approaches denote several techniques for collecting the digital footprints generated by humans when they interact with the techno-social ecosystem.

Data formats range from text, graph, logs, streams, structured, and semi-structured data. Acquisition sources range from social networks, web logs, web content, links, transactions, search engines, social media etc. to the wide variety of smart devices.

Analytical crawling is an example of social sensing. These kinds of methods combine semantic enrichment and machine reading techniques to the aim of gathering semantic-rich data linked across diverse web resources. Crawling techniques have been designed to support search tasks and research has mainly focused on improving the indexing structure, both in efficiency and efficacy.

The new challenge is to push part of the intelligence of social sensing into the gathering process and make it capable of constructing, maintaining and using multi-dimensional network of semantic objects. An example is the study of relationships of important persons from around the world. This might be the network constructed by crawling their co-occurrence on the same article (news, blogs etc.) bound by positive or negative terms, phrases, or opinions [14].

Participatory sensing is another approach to directly involve people as active/passive actors in data collection, by developing tools to elicit facts, opinions and judgements from crowds. Participatory sensing aims at building, on demand, a sensor network infrastructure based on people collaboration towards a common sensing task. The idea of participatory sensing is conceived on the concept that aware participation in collective sensing campaigns enable the creation of social data of higher quality and realism, capable of portraying concurrently different dimensions of social life.

The ability of combining data sensed both by opportunistic and participatory methods is expected to open new avenues for learning systems, capable of recognizing and mapping human behaviour at global and local scale [15,16].

Social mining: provides the analytical methods and algorithms needed to understand human behaviour by means of automated discovery of the underlying patterns, rules and profiles from the massive datasets of human activity records produced by social sensing.

The emergence of big data of human activities, their networked format, their heterogeneity and semantic richness, their magnitude and their dynamical nature pose new scientific challenges.

Social network analysis is an example of social mining approach. The network analysis refers to the study of inter-

[11] The description of these initiatives (called Tuscan Big Data Challenge) is available at the following link: http://www.sobigdata.eu/blog/tuscan-big-data-challenge-20172018.

[12] The number of available resources is growing up thanks to the collaboration between the original partners, new organizations and users.

personal relationships with the purpose of understanding the structure and the dynamics of the fabric of human society.

Social media mining methods address heterogeneous data, especially text, sensed from different on line sources (tweets, mails, blogs, web pages) to the purpose of extracting the hidden semantics from them.

Finally, *understanding human mobility* has emerged as a vital field, leveraging the spatio-temporal dimensions of big data such as mobile phone call records and GPS tracks, to the purpose of mining mobility patterns, daily activity patterns, geographic patterns. Computer scientists put forward mobility data mining, aimed at discovering patterns of interesting mobility behaviours of groups of moving objects. Examples of this process are the mining of frequent spatio-temporal patterns, trajectory clustering, and prediction of future movements.

A recent trend in research focuses on extracting mobility profiles of individuals from their traces, trying to address one of the fundamental and traditional question in the social science: how human allocate time to different activities as part of the spatial temporal socio-economic system, and how groups or individuals interact with different places, at different time of the day in the city [17–19].

2.5 Legal and ethical framework

Data science creates opportunities but also new risks. The use of data science techniques could expose sensitive traits of individual persons and invade their privacy. In particular, social mining approaches require the access to digital records of personal activities that contain potentially sensitive information. These records include, but are not limited to: user mobility data, web page data, social media records and transaction records.

For these reasons, SoBigData takes care of the definition of a legal and ethical framework in accordance with European and national values and legislations. Our RI provides a comprehensive overview on the definition of a framework for processing personal data (and sensitive information) within a research infrastructure. The legal and ethical requirements define the responsibilities of the actors with respect to the respective applicable data protection laws [22,23].

Moreover, SoBigData promotes and adopts ethically grounded collection, management and analysis of this data. SoBigData uses privacy enhancing tools and works towards ensuring its analyses are also just, fair and non-discriminatory and the European data protection law is the focus for the development of a legal and ethical framework inside SoBigData.[13]

[13] From May 2018 the new General Data Protection Regulation (GDPR) shall apply replacing the Data Protection Directive (DPD) and its national implementations.

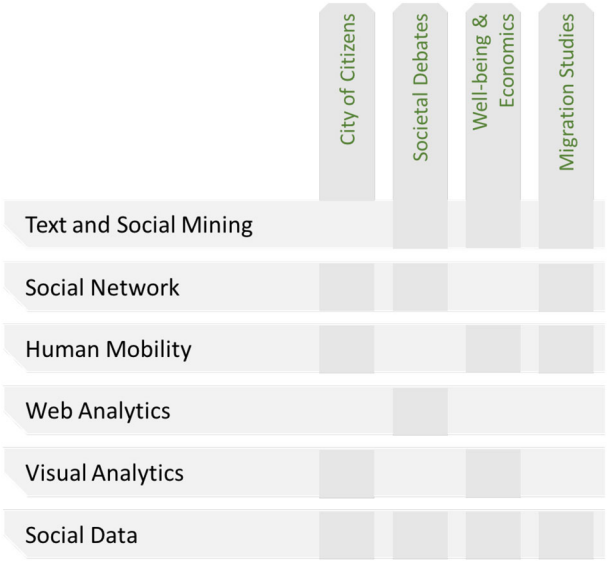

Fig. 3 The relation between thematic areas (listed horizontally) and exploratories (listed vertically)

Summarizing, we can state that this SoBigData component: (i) monitors the compliance of experiments and research protocols with ethical and juridical standards within the framework; (ii) develops big data analytics and social mining tools with value-sensitive design and privacy-by-design methodologies; (iii) boosts excellence and international competitiveness of Europe's big data research in a safe and fair use of big data for research.

3 Data science through SoBigData exploratories

Exploratory projects (or simply Exploratories) are vertical thematic environments built on top of the SoBigData RI. The use of exploratories promotes the effectiveness and usability of the research infrastructure. As shown in Fig. 3, each exploratory covers more than one thematic areas. Exploratory projects are defined by project steering board based on the competencies inside the consortium, the feedback from the advisory board and ideas, needs and feedback from the community. For these reasons, each exploratory is updated continuously, and the number of exploratories continues to grow over time.[14]

Recalling Fig. 1, we can state that an exploratory binds the datasets available in the *big data ecosystem* component with *social mining* methods providing the scientific and research contexts for supporting VA, TA and training, and for promoting networking and dissemination activities. More precisely,

[14] The following link reports list of exploratory available in SoBigData: http://www.sobigdata.eu/exploratories/.

an exploratory serves two purposes: (i) to stimulate internal and external evaluation/feedback of the effectiveness and usability of the RI; (ii) to provide the scientific context of the calls for projects of the TA activities and to promote VA. An exploratory can have sub-lines of research called *stories*. A story represents a specific experiment or use case, and it can be considered a container for binding specific methods, applications, services and datasets. It promotes scientific dissemination, result sharing, and reproducibility.

The definition and the presentation of exploratories are the main actions for supporting and promoting data science. The E-Infra is the engine for providing VA, and the way to enable our users and stakeholders to use our tools remotely. The VREs related to the exploratories are public and are deeply explained in the following sections.

3.1 Migration studies

Big Data are usefull also to understand the migration phenomenon [9]. In this exploratory, the SoBigData RI provides a set of data and tools for trying to answer to some questions about migration flows in Europe and in the world. Through our platform a data scientist studies economic models of migration, can observe how migrants choose their destination countries, and what is the meaning of "opportunities" that a country provides to migrants.

Furthermore, whether there are correlations between number of incoming migrants and opportunities in the host countries [13]. We try to understand how public perception on migration is changing using an opinion mining analysis. Does this perception depend on time and space? Social network analysis enables us to analyse the migrant's social network and discover the structure of the social network for people who decided to start a new life in a different country.

Currently, one story is defined inside this exploratory. It is called *Phases of Migration*, and studies (and defines) the different steps of a migration flow: (i) *the Journey*: at the moment, information about migration flows and stocks comes from official statistics obtained either from national censuses or from the population registries. Given that migration intrinsically involves various nations, data are often inconsistent across data-bases, and offer poor time resolution. With the availability of social big data (e.g. Twitter), it should be possible to estimate flows and stocks from available data in real time, by building models that map observed measures extracted from these unconventional data sources to official data, i.e. nowcasting stocks and flows. In terms of cultural transitions, language mobility is of interest as well.

(ii) *the Stay*: migration generates cultural changes with both long- and short-term effects on the local and incoming population. Migrant integration is generally measured through indicators related to work market integration or social ties (such as mixed marriages). Furthermore, these statistics are available with low resolution and not for all countries. Through SoBigData RI, we observe the integration and perception on migration by analysing big data. For instance, social network sentiment analysis specific to immigration topics allows us to evaluate perception of immigration. Analysis of retail data enables a data scientist to understand whether immigrants are integrated economically but also if they change their habits during their stay. Scientific data help us understand how migration benefits both the host countries and the migrants themselves. Through social network analysis, we can derive novel integration indices that take into account online activity. The effect of multiculturality on overall sentiment is also being observed.

(iii) *the Return*: besides effects on the receiving communities, the source communities may also see effects of migration. One possible scenario is migrants returning to their home countries. SoBigData RI has supported the project "Demal Te Niew", also financed by European Journalism Centre. The project concentrates on studying the returning migrations between Italy and Senegal. The research is based on data journalism, combining data analysis with journalism, and resulted in a documentary. This tells the story of 4 people from Senegal who emigrated to Italy and after several years decided to return to Senegal and start their own business. The documentary was featured in Espresso[15] and El Pais,[16] and presented at the Ethnographic Film Festival, Amsterdam and the International Day of Migrants, Dakar.

Figure 4 shows the workflow related to this story. All the datasets and methods reported in the figure are available through the SoBigData RI. Several studies are ongoing, including developing economic models of migration, nowcasting migration stocks and flows, identifying perception of migration and effect on the leaving and the receiving communities.

3.2 Well-being and economy

The goal of this exploratory is to test the hypothesis that well-being is correlated to business performance of companies. By developing effective policies to reduce internal risks, i.e. operational risks such as credit risk, and external risks (such as systematic and operational risk [12]) of companies, well-being can be systematically improved [10]. The core hypothesis is decomposed into the following more specific sub-hypotheses: (i) there are curves of age- and gender-based segregation distribution in boards of companies, which are characteristic to mean credit risk of companies in a region; (ii) low mean credit risk of companies in a region has strong positive correlation to well-being; (iii) systemic risk correlates

[15] Espresso is an Italian newspaper edited by Gruppo Editoriale l'Espresso.

[16] An important Spanish newspaper edited by PRISA.

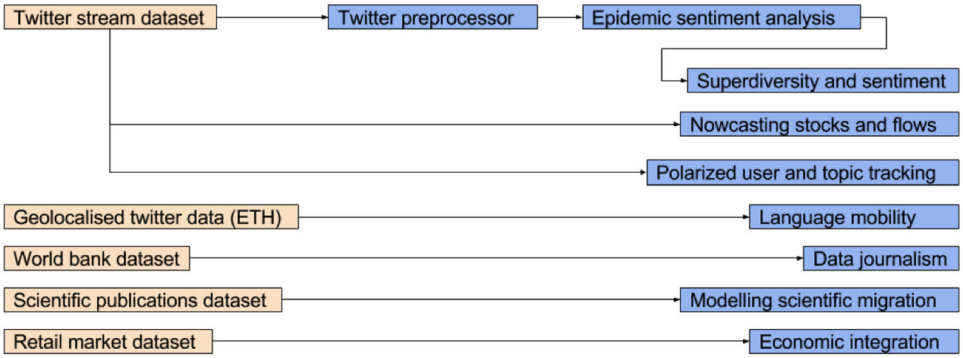

Fig. 4 The workflow of the Migration exploratory: on the left the data sources; on the right the methods applied with their dependencies

highly with well-being indices at national level. The final aim of this exploratory is to provide a set of guidelines to national governments, methods and indices for decision-making on regulations affecting companies in order to improve well-being in the country.

Instead of using a workflow as the one proposed for Migration Studies, SoBigData RI supports well-being data science proposing a set of experiments documented in the form of walk-throughs. The latter guide the user step-by-step in running the experiments.[17] The experiments demonstrate the benefits of the SoBigData RI from the following angles: (i) to support pan-European studies (e.g. experiments to study a phenomenon in Italy complemented with similar studies with corresponding Estonian data); (ii) to support regional studies (e.g. developing models for nowcasting regional GDP); (iii) to support analysis of a phenomena with a combination of longitudinal and snapshot data; (iv) to access and use public and proprietary datasets; (v) to use existing methods and tools for advanced sociological studies.

In this context, let us consider the hypothetical case of a national government that has to make a decision on whether to enforce a regulation to control which demographic properties, e.g. gender-based distribution, company boards should comply to improve well-being in the country. As part of the decision-making procedure, a team of social scientists has been hired to study related phenomena. The following data are available:

- *CDR data*: they consist of telecommunication records. Each data entry contains the following information about a single call: the phone numbers of the subscribers originating and receiving the call, the starting time of the call, the call duration, the identification of the cell from which the call started, the latitude and longitude of the cell from which the call started, as well as the call type.
- *Retail dataset*: it consists of customer transactions. Each transaction contains a complete set of information about the customer and the items purchased.
- *Company dataset*: it contains various metadata about companies. Specifically, it includes a company identifier, legal type, share capital, status, number of workers, registration date and registration district, physical address, and activity code.
- *Board member data:* this dataset contains information for company board members. The data include a board member's birth data, country of residence, gender, affiliated company, as well as board membership role and duration.
- *Balance sheet + asset class capitalization dataset*: it contains the equity and exposure to twenty asset classes for a large number of banks from 2001 to 2014. More than ten thousand (10,000) banks appear in the dataset at least once.
- *Stock market transaction data*: this dataset contains stock market data from Thomson Reuters, covering the years from 2006 to 2014. Each stock transaction contains the following information: equity ticker, date, time of transaction, GMT offset, type, price, volume, bid price, ask price (if quote), execution time.

The data analysis for this exploratory includes the following set of methods:

- *Entropy measures for poverty*: this method is related to human mobility with socio-economic indicators. Its input consists of CDR data, on the one hand, as well as measures for poverty on the other. Its output consists of the radius of gyration and mobility entropy aggregated per area unit, as well as the correlation between mobility and the poverty measures.
- *Nowcasting GDP from retail market data*: this method tests nowcast GDP using retail market data and provides estimates of well-being by exploiting human shopping behaviour. The analysis makes use of the bipartite networks of customers and products for different time windows and market categories. Its input consists of retail data and GDP indicators. Its output consists of a product sophistication (SOP) index and an estimate for economic well-being.

[17] The walk-throughs are currently freely available and published at http://sobigdata.ee.

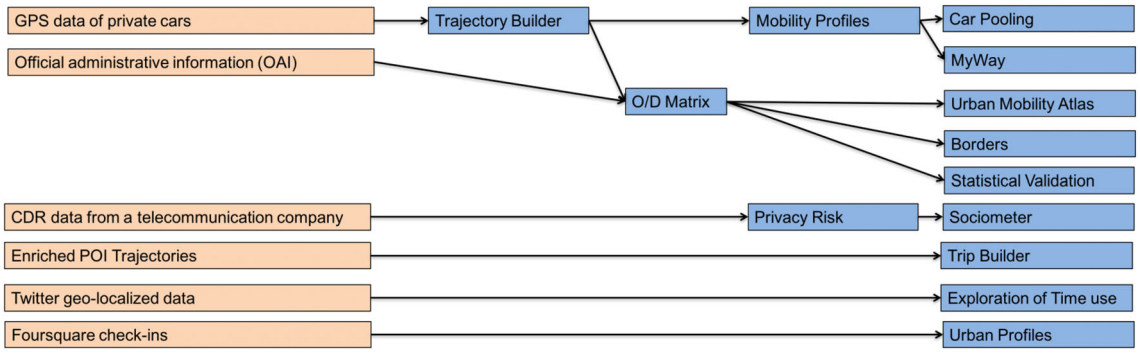

Fig. 5 The workflow of City of Citizens story: on the left the data sources; on the right the methods with their dependencies

- *Systemic risk estimation*: it consists of three sub-methods and its goal is to assess the contribution of individual nodes to the distress risk of a network. Its input consists of balance sheet data and stock market transaction data. Its output consists of a systemic risk indicator for each node of the network [2].
- *Credit risk scoring*: this method evaluates the credit risks of companies. The input consists of company data, board member data, segregation indices, and systemic risk indices. It provides a model that produces the probability of default in X days. The method employs the random-forest machine learning algorithm to produce its predictions.
- *Segregation discovery*: this method provides exploratory analysis of segregation indexes of social groups in company boards. Its input consists of company data and board member data. Its output is an OLAP data cube.
- *Reconstructing financial networks*: the goal of this approach is to build financial networks. Its input consists of data regarding bank assets and liabilities, as well as the total number of links one aims to discover. Result is a network graphs annotated with topological indicators.

3.3 City of citizens

The objective of this exploratory is to collect stories related to geo-referenced data that describe a city, a territory, or an entire region. There are several studies and different methods that employ a broad variety of data sources to build models about mobility of people and city characteristics in the scientific literature [7,8].

This exploratory addresses the needs of three different kinds of users: *researchers*, who search for datasets, models, or methods, evaluate and compare existing algorithms, or publish new solutions; *data analysts*, who search for existing methods and methodologies to be applied in different contexts and evaluate the results and *administrations*, who search for the results of methods that have been applied to datasets of interest, or look for collaborations with the previous categories.

Currently, this exploratory contains one story. The story focuses on studying how people move within a city, considering different categories of citizens and places. This story reports the analysis of two cities: Pisa and Florence. These two cities are selected for the availability of geo-located data sources. The following datasets are considered, and registered in the catalogue of the SoBigData RI:

- *GPS traces*: a dataset of private vehicles crossing Tuscany. The number of records is about 11 millions of trips; number of users is 150,000.
- *Call data records* (CDRs): the dataset contains mobile phone records collected in the provinces of Pisa, Lucca, Livorno and Firenze in March 2014. It contains about 50 millions of CDRs, and the antennas' coverage. Number of CDRs is about 50 million; number of users is about 860,000; number of antennas is about 450.
- *Foursquare* check-ins: about 15 million tweets that point to public Foursquare check-ins.
- *Flickr* geo-localized photos: the dataset contains a knowledge base built with data coming from Flickr and Wikipedia. It covers three Italian cities: Rome, Florence, and Pisa. These are three of the most important cities from a sightseeing point of view and, thus, the dataset guarantees variety and diversity in terms of size and richness of available content.
- *Official administrative information*: a set of geometries describing census sectors with some demographic information. Number of sectors is about 20,000.
- *Geo-localized twitter data*: a dataset of geo-localized Tweets in the area of Tuscany in a period of one year.

The results of this story are useful for both local administrators and citizens. The local administrators have a tool to quantify accurately city's traffic and understand city's usage. In this perspective, they could take better decisions to manage mobility.

Citizens take information for checking the traffic situation in real time and they could choose the best and fastest way or take advantage of a car-pooling service. Additionally, the methods to build models and extract analytical results over the datasets are shown in the workflow[18] of Fig. 5.

3.4 Societal debates

Social media sites are becoming a crucial component in the public sphere, and discussions hosted on them reflect public perception of a diverse set of issues, such as security, climate, environment, migration, culture, athletics. By analysing discussions on social media, we get a better understanding of how certain matters become issues of concern for the public, and how the discussion and controversies around them evolve [11].

In this context, a data scientist can analyse how citizens debate contentious topics in social media, including as examples the British in/out EU referendum and other opportunistic debate datasets provided by SoBigData RI partners. Within this exploratory, we intend to employ social mining methods to answer the following research questions: Who is participating in these public debates? What do debate participants believe? How are debate participants influencing one another? What arguments are being made and by whom? What is the "big picture" response from citizens to a policy, election, referendum or other political event? This kind of analysis allows scientists, policy makers and citizens to understand the online discussion surrounding polarized debates. The personal perception of online discussion on these issues on social media is often biased by the so-called filter bubble, in which automatic curation of content and relationships between users negatively affects the diversity of opinions available to them. Making a complete analysis of online polarized debates enables the citizens to be better informed and prepared for political outcomes.

A story inside this exploratory includes a complete study on a hot topic about Brexit. This case study employs a collection of tweets related to British EU membership referendum. Tweets were gathered according to the presence of a hashtag related to the debate, and as of the 17th of June, half a million tweets were collected per day. The tweets were processed in a streaming fashion using a GATE-based debate analysis pipeline, designed to identify topics, sentiment, named entities, and vote intent.[19] The GATE real-time analytics infrastructure (integrated into SoBigData RI) processes tweets faster than they were arriving via the Twitter Streaming API. As part of this story, we also extended the pipeline to identify the location at which tweets were posted as an NUTS geographical region, and to classify the authors of each tweet according to the kind of entity they represent (celebrity, journalist, member of the public, organization, place). After processing, the tweets were stored in a Mímir[20] index. Mímir stores entities, metadata and other derived features from the metadata of the tweets and supports complex search queries, combining full-text and semantic search over tweets, e.g. "What are leave voters in South Yorkshire saying about immigration?". Visualizations based on such queries are generated using Prospector, which supports tag clouds, co-occurrence matrices and association diagrams [3,4].

Task-specific visualizations have also been created, and integrated into a public-facing web page for exploring the debate. Tag clouds are generated both for the most popular hashtags and the most popular discussion topics in all tweets, tweets indicating an intention to vote leave and tweets indicating an intention to vote remain. Clicking on elements in these visualizations allows relevant tweets to be viewed in order to carry out more in-depth analysis (Fig. 6).

4 Conclusions

Big data managing and exploitation require a complete re-design of existing architectures and propose new challenges on data management, privacy, and scalability. Once again, the sheer size and the complexity of big data call for novel analytical methods. The kind of measures provided by the data and the population sample they describe cannot be easily modelled through standard statistical frameworks, which therefore need to be extended to capture the way the data are generated and collected. In this perspective, the challenge is particularly tough: which data mining tools are needed to master the complex dynamics of people in motion and construct concise and useful abstractions out of large volumes of data is, by large, an unanswered question.

Based on these premises, this work has shown why the problem of analysing big data is relevant for our society, and why the availability of a distribute research infrastructure such as SoBigData is fundamental to support and promote

[18] SoBigData project is developing a language and an execution platform for representing scientific process in highly heterogeneous e-Infrastructures in terms of so-called hybrid workflows. Currently, SoBigData workflows can express sequences of manually executable actions, which offer a formal and high-level description of a reasoning, protocol, or procedure, and machine-executable actions, which enable the fully automated execution of one (or more) web services [20].

[19] The GATE platform provides end-to-end text processing solutions. A last version of the GATE platform is available at cloud.gate.ac.uk.

[20] Mímir is a DBMS used by GATE Infrastructure for collecting documents with information stored as annotations.

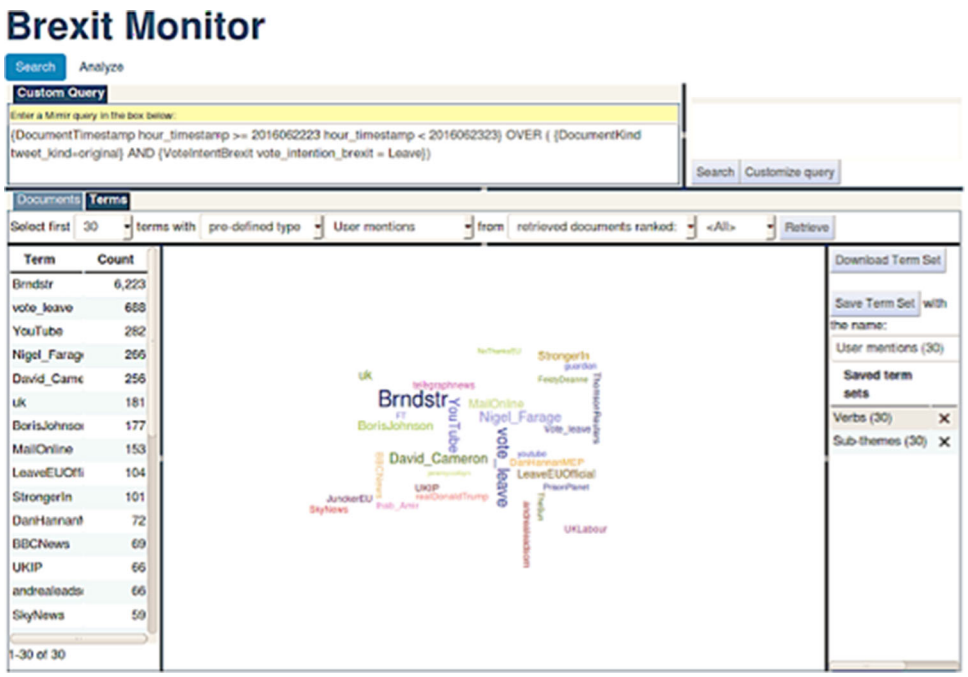

Fig. 6 A screenshot of the Brexit Monitor tool. A demo of the Brexit Monitor is available at the following link: http://demos.gate.ac.uk/sobigdata/brexit

data science. SoBigData RI provides *appropriate analytical* technologies 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. We introduced a set of case studies that are useful by a data scientist in order to perform her/his own studies and researches. We highlighted the key role of the data in a research context especially when we consider social mining. For these reasons, we promoted a data science methodology based on interdisciplinary approaches called exploratories. Furthermore, we have stated why legal and ethical aspects must be a pillar for the definition of a RI supporting data science and using VRE tools we can guarantee open data and reproducibility.

Another substantial part of the SoBigData RI is dedicated to networking activities, in order to create a world-wide user community around the infrastructure, as well as attract new contributors of datasets and methods, which they can integrate into the SoBigData RI. A user can join to the SoBigData RI by compiling a simple registration form and can request to become member of the VREs available inside the platform. A registered user can access to the RI catalogue, can execute an available method or get a public dataset. Furthermore, the user can add a method or/and a dataset making them available to the SoBigData community, as well as can run the new integrated method using RI computational facilities. The set of SoBigData facilities promote collaboration with other users in an innovative and effective way, e.g. enabling them to post or be notified of news about the availability of a new resource, to post news about planned experiments or ask for expert support and comments on a given idea or issue.[21]

Our effort for the future is to develop more advanced (and easy-to-use) tools for uploading and, sharing datasets, for integrating and invoking methods and submitting innovative ideas for implementing new experiments. Finally, we are integrating the platform with a workflow engine capable of representing scientific processes in highly heterogeneous SoBigData RI in terms of "hybrid workflows", that can express sequences of human-executable actions, i.e. formal descriptions guiding users to repeat a reasoning, protocol, or manual procedure, and machine-executable actions, i.e. encoding of the automated execution of one (or more) web services.

Acknowledgements This work has been supported by the *EC H2020 INFRAIA-1-2014-2015*—Project "SoBigData Research Infrastructure: Social Mining & Big Data Ecosystem" (Grant Agreement No. 654024). This work would not have been possible without the contributions of all the people involved in the SoBigData project.

References

1. Geraci, A.: IEEE Standard Computer Dictionary: Compilation of IEEE Standard Computer Glossaries. IEEE Press, Piscataway (1991)
2. Greenwood, R., Augustin Landier, A., Thesmar, D.: Vulnerable banks. J. Financ. Econ. **115**(3), 471–485 (2015)

[21] Using SoBigData Gateway the user can access all the information required for register and upload a dataset or a method (https://sobigdata.d4science.org/group/sobigdata-gateway).

3. Maynard, D., Greenwood, M., Roberts, I., Windsor, G., Bontcheva, K.: Real-time social media analytics through semantic annotation and linked open data. In: Proc of 2015 ACM Web Science, Oxford, United Kingdom (Jul 2015)
4. Maynard, D., Bontcheva, K.: Understanding climate change tweets: an open-source toolkit for social media analysis. In: Proc. of EnviroInfo 2015, Copenhagen (Sep. 2015)
5. Candela, L., Castelli, D., Pagano, P.: Virtual research environments: an overview and a research agenda. Data Sci. J. **12**, GRDI75–GRDI81 (2013). https://doi.org/10.2481/dsj.GRDI-013
6. Candela, L., Castelli, D., Manzi, A., Pagano, P.: Realising virtual research environments by hybrid cata infrastructures: the D4Science experience. In: International Symposium on Grids and Clouds (ISGC), Proceedings of Science PoS(ISGC2014) (2014)
7. Trasarti, R., Guidotti, R., Monreale, A., Giannotti, F.: MyWay: location prediction via mobility profiling. Inf. Syst. **64**, 350–367 (2017). https://doi.org/10.1016/j.is.2015.11.002
8. Nanni, M., Trasarti, R., Monreale, A., Grossi, V., Pedreschi, D.: Driving profiles computation and monitoring for car insurance (CRM). ACM TIST **8**(1), 14:1–14:26 (2016). https://doi.org/10.1145/2912148
9. Moise, I., Gaere, E., Merz, R., Koch, S., Pournaras, E.: Tracking language mobility in the twitter landscape. In: (IEEE) International Conference on Data Mining Workshops (ICDM) Workshops (2016). https://doi.org/10.1109/ICDMW.2016.0099
10. Mazzarisi, P., Lillo, F.: Methods for Reconstructing Interbank Networks from Limited Information: A Comparison. In: Abergel, F., et al. (eds.) Econophysics and Sociophysics: Recent Progress and Future Directions. New Economic Windows. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-47705-3_15
11. Garimella, K., De Francisci Morales, G., Gionis, A., Mathioudakis, M.: Reducing controversy by connecting opposing views. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM '17 (2017). https://doi.org/10.1145/3018661.3018703
12. Grossi, V., Romei, R., Ruggieri, S.: A case study in sequential pattern mining for IT-operational risk, machine learning and knowledge discovery in databases. In: European Conference (ECML/PKDD), pp 424–439 (2008)
13. Coletto, M., Esuli, A., Lucchese, C., Muntean, C., Nardini, F.M., Perego, R., Renso, C.: Sentiment-enhanced multidimensional analysis of online social networks: perception of the mediterranean refugees crisis. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1270–1277 (2016)
14. Bontcheva, K., Rout, D.P.: Making sense of social media streams through semantics: a survey. Seman. Web **5**(5), 373–403 (2014)
15. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., Tesconi, M.: The paradigm-shift of social spambots: evidence, theories, and tools for the arms race. CoRR abs/1701.03017 (2017)
16. Cresci, S., Tesconi, M., Cimino, A., Dell'Orletta, F.: A linguistically-driven approach to cross-event damage assessment of natural disasters from social media messages. WWW (Companion Volume), 1195–1200 (2015)
17. Trasarti, R., Guidotti, R., Monreale, A., Giannotti, F.: MyWay: location prediction via mobility profiling. Inf. Syst. **64**, 350–367 (2017)
18. Nanni, M., Trasarti, R., Monreale, A., Grossi, V., Pedreschi, D.: Driving profiles computation and monitoring for car insurance CRM. ACM TIST **8**(1), 14:1–14:26 (2016)
19. Guidotti, R., Trasarti, R., Nanni, M., Giannotti, F.: Towards user-centric data management: individual mobility analytics for collective services, pp. 80–83. MobiGIS (2015)
20. Candela, L., Manghi, P., Giannotti, F., Grossi, V., Trasarti, R.: HyWare: a HYbrid Workflow lAnguage for Research E-infrastructures, D-Lib Magazine **23**(1/2) (2017)
21. Grossi, V., Rapisarda, B., Romano, V.: Fact sheets aimed at different stakeholders, SoBigData project deliverable. https://goo.gl/aCC8Le (2015)
22. Hänold, S., Forgó, N., van den Hoven, J., Mahieu, R., van Putten, D.: Legal and ethical framework for SoBigData 1, SoBigData project deliverable. https://goo.gl/NUiWhR (2016)
23. Hänold, S., Forgó, N., van den Hoven, J., Mahieu, R., van Putten, D., Lishchuck, I.: Legal and ethical framework for SoBigData 2, SoBigData project deliverable. https://goo.gl/5MLkzN (2017)