Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
iCity & Big Data—Perspective

Big Data Research in Italy: A Perspective

Sonia Bergamaschi[a], Emanuele Carlini[b], Michelangelo Ceci[c], Barbara Furletti[d], Fosca Giannotti[d], Donato Malerba[c,e,*], Mario Mezzanzanica[f], Anna Monreale[d], Gabriella Pasi[g,*], Dino Pedreschi[d,h], Raffele Perego[b], Salvatore Ruggieri[h]

[a] Department of Engineering "Enzo Ferrari," University of Modena and Reggio Emilia, Modena 41125, Italy
[b] High Performance Computing Laboratory, Institute of Information Science and Technologies of the Italian National Research Council (ISTI-CNR), Pisa 56124, Italy
[c] Department of Computer Science, University of Bari Aldo Moro, Bari 70125, Italy
[d] Knowledge Discovery and Data Mining Laboratory, ISTI-CNR, Pisa 56127, Italy
[e] Big Data Laboratory, National Interuniversity Consortium for Informatics, Rome 00185, Italy
[f] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan 20126, Italy
[g] Department of Computer Science, Systems and Communications, University of Milano-Bicocca, Milan 20126, Italy
[h] Department of Computer Science, University of Pisa, Pisa 56127, Italy

ARTICLE INFO

Article history:
Received 16 December 2015
Revised 4 June 2016
Accepted 13 June 2016
Available online 30 June 2016

Keywords:
Big data
Smart cities
Energy
Job offers
Privacy

ABSTRACT

The aim of this article is to synthetically describe the research projects that a selection of Italian universities is undertaking in the context of big data. Far from being exhaustive, this article has the objective of offering a sample of distinct applications that address the issue of managing huge amounts of data in Italy, collected in relation to diverse domains.

1. Introduction

In the last few years, initiatives, events, and projects related to big data have proliferated, both in research centers/academies and in industry. The daily production of huge quantities of data related to various and diversified aspects of social life (including mobile phone data, social data, city related data, Web-based data, and health-related data) offers an unprecedented opportunity to observe and learn about peoples' preferences and behavior and to exploit this information in order to improve certain aspects of peoples' lives.

In response to this disruptive change—a change that opens up new economic perspectives—the European Commission has called on national governments to wake up to the "big data" revolution[†]. The European digital economy has indeed been slow in embracing the data revolution compared with the US, and also lacks comparable industrial capability. To recover from this delay, considerable funding has been and will be provided by the European Commission as well as European countries to support research and innovation actions related to value generation from big data. To properly address this objective, various issues must be taken into consideration, ranging from the definition of powerful and technologically suited infrastructures to support-intensive data-driven computations (on both hardware and software) to the setting-up of multidisciplinary teams to properly and fruitfully extract knowledge from data in var-

[*] Corresponding authors.
E-mail addresses: donato.malerba@uniba.it; pasi@disco.unimib.it

[†] "Towards a thriving data-driven economy," communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee, and the Committee of the Regions, Brussels, 2 Jul 2014.

http://dx.doi.org/10.1016/J.ENG.2016.02.011

ious domains.

Despite this delay, the European big data market controls the second largest market share, at 20% in terms of revenue in the global big data market [1]. Germany, the United Kingdom, France, and Italy are key countries in this market. In particular, the Italian big data market has grown rapidly in the last year, and significant investments are expected shortly from both the private and public sectors. This short survey reports some of the applications and projects that are undertaken by Italian universities with respect to the challenge of big data; in particular, it reports projects related to improving citizens' lives. As interesting examples of applications of technologies related to big data management, Section 4 describes a system aimed at monitoring the production and consumption of energy, while Section 5 synthetically presents a prototype aimed at analyzing Web job vacancies collected from five European Union (EU) countries and extracting the requested skills from the data.

It is important to note that this survey is far from being exhaustive, both with respect to active research groups on these topics (since such groups are considerably more numerous than those referred to in this paper) and with respect to projects, which in Italy are much more numerous than those reported here. The main aim is to offer the reader a flavor of the problems that academies are addressing with respect to this important issue.

A recent national initiative on big data is represented by the CINI "Big Data" Laboratory. CINI (www.consorzio-cini.it) is the Italian National Interuniversity Consortium for Informatics, a consortium of 41 public Italian universities that promotes and coordinates scientific activities of research and technological transfer, both theoretical and applicative, in several fields of computer science and computer engineering. The consortium is a founding member of the Big Data Value Association (www.bdva.eu), the industry-led contractual counterpart to the European Commission for the outlining and implementation of the European strategic research agenda on big data. In addition, the CINI "Big Data" Laboratory—which focuses on data that is distributed throughout the whole national territory—has the aim of being a center of Italian expertise for the development of knowledge and technologies in the fields of big data and data science. Thirty-three Italian universities and about 300 researchers currently adhere to this initiative.

The next sections briefly present a few projects carried out by different Italian universities and research centers that address the issue of big data and that aim to improve various aspects of people's lives. These projects are related to distinct applicative domains, including understanding city dynamics, the Italian healthcare system, forecasting energy production in photovoltaic power plants, and managing job offers. The last two sections address the important issues of privacy and big data usability.

2. Understanding human and city dynamics with mobile phone data

Cities have always been complex systems of people, things, environments, and activities, and their rapid evolution has led to an unavoidable increase of complexity. This fact is pushing scientists to leave traditional paradigms of model-driven analysis in favor of data-driven approaches, opening the era of big data analytics. Digital signs that people produce every day by interacting with devices, social media, and other technological systems give unprecedented opportunities to study and understand city dynamics and social behavior from several perspectives. Understanding these dynamics means being able to anticipate the impact of phenomena and to support policies and planners in responding to citizens' needs.

Mobile phone data actually represent a proxy for studying and measuring cities and citizens, allowing us to identify peoples' presence at the urban level [2–4], to reconstruct their mobility [5–8] and sociality [9], and to study the impact of events in cities [4,10].

2.1. Mobile phones and origin and destination (OD) matrix estimation

The estimation of presences and flows between preferred locations can be used to reconstruct an origin and destination (OD) matrix [5,6] that is useful for inferring transport demand models and for understanding infrastructure requirements. In Ref. [6], a long-term analysis of individual call traces is performed, in order to reconstruct systematic movements (i.e., movements with a high frequency) between the two most significant locations for an individual. Such locations, typically associated with home and work, are identified among the locations from which an individual made the largest number of calls. After having identified the systematic movements between these locations, the OD matrix summarizes the expected traffic flows between spatial regions.

2.2. Mobile phones for novel demography and city user estimation

The possibility of measuring and monitoring social phenomena has increased the interest in the use of big data to support official statistics [5]. Since administrative data cannot be collected with high frequency and often do not contain accurate information on mobility, calling data are being used more and more to integrate traditional data sources, allowing the construction, for example, of permanent observatories of the cities [3] and the identification of actual types of city users. In Refs. [2] and [4], the Sociometer, an analytical framework aimed at classifying mobile phone users into behavioral categories, is presented. The analytical process starts with the construction of spatio-temporal profiles synthesizing the presence of the individuals in the area of interest. Then, by applying a data mining method, different people categories are learned, and annotated profiles belonging to residents, dynamic residents, commuters, and visitors are produced. In Ref. [5], starting from the result of the Sociometer, an OD matrix at the municipality level is created in order to observe the inter-city mobility of the individuals. By producing statistics comparable to those obtained by the National Institute of Statistics (Italy), a safe way is offered to integrate existing population and flow statistics with the continuously up-to-date estimates obtained from mobile phone data.

2.3. Mobile phones, mobility diversity, and economic development

Studies become more and more challenging when there is a need to investigate society status in order to improve living conditions. In Ref. [8], starting from nation-wide mobile phone data, the authors extract a measure of mobility diversity and mobility volume for each individual, and investigate the correlations with external socioeconomic indicators. Diversity is defined in terms of the entropy of the individual users' trajectories, while volume of the mobility is measured by the characteristic distance traveled by an individual. The experiments show that mobility is correlated with wellbeing indicators (such as education level, unemployment rate, income, and deprivation), demonstrating a high predictive power of mobility behaviors with respect to the socioeconomic development of cities. In another exploration of the social dimension, an interesting result emerges in Ref. [9] from comparing mobility with the social network extracted from calls. The similarity in the movements and proximity in the social network appears to be strongly related, leading to the conclusion that people not connected in the network, but topologically close and with similar mobility patterns, are likely to share a network connection in the future.

2.4. Mobile phones and big event estimation

The possibility of monitoring and registering peoples' reactions to events in terms of displacement is of great interest for public administrations [10]. A similar consideration can be provided for the impact of events in cities [4], in order to design adequate plans for security and mobility. Ref. [10] presents correlation pattern analysis, a process for extracting interconnections between different areas caused by events in the city. By analyzing the density of calls at a collective level, the presence of people is estimated, and significant co-variations of presences is derived by using time- and space-constrained sequential pattern analysis. Ref. [4] presents a method for measuring the impact of events (e.g., festivals, music and artistic shows, and seasonal events) at the urban level by exploiting the people-profiling supplied by the Sociometer. The variation of city users' composition in the area of interest and over a specific period of time is analyzed using statistical methods and a multi-classification analysis. The multi-classification aspect allows us to investigate how the composition of the population changes when moving the analysis from small areas (e.g., the city's historical center) to larger ones (e.g., the suburbs). Experiments confirm the capability of the Sociometer to identify people composition at the urban level and the validity of the whole method in measuring the impact of big events in both small and big cities.

2.5. Mobile phone models of human mobility

Thanks to the ubiquitous nature and diffusion of these data, new characterizations of human dynamics have been possible. Human mobility, approximated for decades with random walk or Lévy flight, has instead revealed a high degree of temporal and spatial regularity that does not exclude heterogeneity in the patterns. Through a study of the radius of gyration computed on mobile phone trajectories, it has been discovered that people spend most of their time in a small number of locations. This result has allowed scientists to investigate mobility deeper, finding that the considerable variability in the characteristic traveled distance of individuals coexists with a high degree of predictability in their future locations. This apparent contradiction has been explained by further analyzing the impact of systematic movements, finding that two new categories of travelers exist: the returners and the explorers. The systematic mobility of the returners is estimated by their radius of gyration and characterized by recurrent movements between a few preferred locations. The explorers tend instead to move between a larger number of different locations and their systematic mobility gives only a small contribution toward their overall mobility [7].

3. A big data case for Italian healthcare

Standard healthcare practices are becoming progressively based on evidence of medical knowledge extracted from large volumes of medical data. In all developed countries, healthcare providers collect and manage large amounts of complex, heterogeneous data. This huge availability of data promises virtually infinite possibilities in the continuous process of improving the sustainability and quality of healthcare systems, ranging from personalized medicine, disease prevention, and effective healthcare organization [11]. However, a large number of patients receive healthcare from different providers, thus creating a fragmentation of e-health data spread among many organizations. The integration and harmonization of these data is thus becoming increasingly important.

In this context, the Italian tax-based, public healthcare system presents distinctive challenges, due to its universal coverage and regional administration. Italy's population is among the oldest in the world, and effective management of chronic conditions [12] is of paramount importance for the patients, in order to prevent complications and disability, and for the nation to ensure economic sustainability.

The organization of the Italian healthcare system is hierarchical and decentralized. The national level is responsible for ensuring the general objectives and fundamental principles of the healthcare system. On the other hand, regional governments (21 in total) are responsible for delivering healthcare through a network of local healthcare units (LHUs, 10 per region on average). Due to this decentralized and independent organization, healthcare data management systems are not interoperable. In this context, National Agency for Regional Healthcare Services (AGENAS) (the national agency for the coordination of regional healthcare systems), in collaboration with Regional Health Agency (ARS) of Tuscany and the National Research Council (Italy), is developing a big data analytics platform aimed at providing unified analytics tools on administrative e-health records managed by regional units.

The THEMATRIX platform supports the whole big data analytics life cycle, from distributed data acquisition/storage to the design and parallel deployment of analytics and the presentation of results. It allows the hiding of the extreme diversities of regional information systems by supporting extraction and remapping to a common schema of administrative records by recording all the interactions of every citizen with the public healthcare system.

The challenges stemming from the data collection regard the diversity and heterogeneity of both data models and data storage technologies exploited at the regional levels. Although the data model is common at the national level, there is little or no enforcement of this model at the local levels. In addition, LHUs have complete freedom in the choice of data management technologies, causing a proliferation of solutions for data storage services and access interfaces (ranging from open-source installations to full-blown enterprise databases). The THEMATRIX platform, which is under development, gathers long-term data from LHUs and organizes them in a common format that can be exploited in integrated studies. An important aspect of data collection is data anonymity. Indeed, when e-health data are managed, privacy is one of the most compelling concerns [11]. Our data collection mechanisms anonymize patients' records according to the guidelines established by the national authority of privacy. Data obfuscation allows very useful cross-regional analysis to be performed, while hiding the identities of the patients. This process is performed at the local level and enforced at the national one, where any personally identifiable information is hidden without hindering the value of the national-level analysis performed.

The data analysis interface provides the epidemiologist with flexible access to data and with a graphical interface for the definition of rule-based algorithms for knowledge extraction. Data transformation and analytics exploit a flexible domain-specific language that provides the possibility to conduct both intra-regional and nation-wide studies of patterns, causes, and effects of health and disease conditions in the respective populations. The programmable computational engine organizes the computation as a directed acyclic graph (DAG), in which every node represents a task to be applied to the stream of patient records.

Preliminary studies conducted on selected pilots have focused on the identification and forecasting of a few chronic conditions, such as diabetes or cardiovascular diseases. The key performance indicators computed on the efficiency and effectiveness of the care of these conditions allow the regional public healthcare systems to be compared on an objective basis, and the quality of forecasting algorithms to be enhanced, due to the huge amount of data available. To date, the platform has been deployed and tested on pilot LHUs spread all over Italy and on two regional agencies. The data available consists of four years of administrative records of about 7 million citizens. In order to enhance predictive models exploiting these healthcare data, the administrative records of about 60 000 patients have been anonymously matched with the specific patients' health conditions, and assessed by the primary care physicians that mediate every relation between the patient and the public healthcare system in Italy. Next year, the project aims at deploying a THEMATRIX analytics solution in at least 10 regions, covering more than half of the Italian population. In order to support this country-level big data analytics challenge, the parallelization of the DAG computation will be enhanced. The requirement is to provide a flexible and effective exploitation of the heterogeneity of LHU hardware, ranging from low-spec commodity machines to large enterprise clusters supporting Apache Spark and Hadoop.

4. Big data for energy

The urgent need to reduce pollution emission has made renewable energy a strategic sector [13], especially for the EU. This has resulted in an increasing presence of renewable energy sources and thus, significant distributed power generation. The main challenges faced by this new energy market are grid integration, load balancing, and energy trading. First, integrating such distributed and renewable power sources into the power grid, while avoiding decreased reliance and distribution losses, is a demanding task. In fact, renewable power sources, such as photovoltaic arrays, are variable and intermittent in their energy output, because the energy produced may also depend on uncontrollable factors, such as weather conditions [14]. Second, the main players in the energy market—the distributors and smaller companies that act between offer (traders) and request in the supply chain—have to face uncertainty not only in the request but also in the offer, while planning the energy supply for their customers. Third, the power produced by each single source (especially from renewable energy) contributes to defining the final clearing price in the daily or hourly market [15], thus making the energy market very competitive and a true maze for outsiders.

In order to face these challenges, it is of paramount importance to monitor the production and consumption of energy, both at the local and global level, to store historical data, and to design new, reliable prediction tools. The Virtual Power Operating Center (Vi-POC) project aims at designing and implementing a prototype that is able to achieve this goal [16,17]. Due to the heterogeneity and the high volume of data, it is necessary to exploit suitable big data analysis techniques, in order to perform an efficient access to data that cannot be obtained with traditional approaches for data management. However, due to the availability of new (low-cost) technologies, small producers are also able to collect data about their business. Indeed, data coming from small production plants are quite heterogeneous: They arrive at a continuous (fast) rate and their volume increases constantly. Moreover, in order to consider uncontrollable factors, such as weather conditions, it is necessary to store weather information (e.g., temperature, humidity, wind speed, etc.), both observed and forecasted, which is collected by querying weather services.

In this perspective, the Vi-POC project has been developed in order to support (renewable) energy providers with a framework for collecting, storing, analyzing, querying, and retrieving data coming from heterogeneous energy production plants (such as photovoltaic, wind, geothermal, Sterling engine, and running water) distributed over a wide territory. Moreover, Vi-POC features an innovative system for the real-time prediction of the energy production that integrates data coming from production plants and weather production services.

In Vi-POC, an HBase storage system has been designed for storing weather information and plant sensor data. These data are exploited by clients running data mining algorithms with the aim of predicting the output power of plants in the next 24/48 hours. Each plant periodically sends all the data collected by the installed sensors. The time granularity is set based on the type and the dimension of the plant. Data coming from plants usually consist of different measures, gathered from several sensors at a given timestamp. Indeed, the number and type of sensors may differ among plants. On the other hand, forecasted data consist of various predicted weather parameters forecasted for a given time and location.

For (renewable) energy power prediction, in the literature, several data mining approaches have been proposed. Researchers typically distinguish between two classes of approaches: physical and statistical. The former relies on the refinement of numerical weather prediction forecasts with physical considerations (e.g., obstacles and orography) [18] or measured data (an approach often referred to as model output statistics or MOS) [19], while the latter is based on models that establish a relationship between historical values and forecasted variables.

Despite the existence of such data mining algorithms applied in renewable energy power forecasting for learning adaptive models [15,20], there is no consensus on the spatio-temporal information to be taken into account, the learning setting to be considered, and the learning algorithms to be used. The dimensions of analysis considered in the predictive models implemented in the Vi-POC framework are:

(1) The consideration of the spatio-temporal autocorrelation [21]: This characterizes geophysical phenomena to obtain more accurate predictions. Spatial autocorrelation is taken into account by resorting to two spatial statistics, namely local indicators of spatial association (LISA) and principal coordinates of neighbour matrices (PCNM), while temporal autocorrelation is considered by exploiting different forms of temporal statistics.

(2) The learning setting to be considered: This is done either by using a simple output prediction for each hour or by using a structured output prediction model (namely, a 24-element vector corresponding to the 24 hours of the next day).

(3) The learning algorithms: The performances of artificial neural networks, which are most often used for photovoltaic production prediction forecasts, are compared to those of regression trees and *k*-nearest neighbors algorithm (or *k*-NN for short, it is implemented in the Apache Spark framework [22]) for learning adaptive models. The results obtained on two datasets show that taking into account spatio-temporal autocorrelation is beneficial.

However, the most important aspect is the learning setting: The structured output prediction setting outperforms by a great margin the non-structured output prediction setting. Finally, the results show that regression trees provide better models than artificial neural networks and *k*-NN prediction models.

5. Job offers and big data

The number of job vacancies advertised through specialized Web labor market portals and services has been growing apace over the last years, enabling new ways for recruitment (also known as e-recruitment) and labor market analyses (also known as labor market intelligence). Informally speaking, a Web job vacancy can be seen as raw text posted several times on different Web sources, specifying ① the job title, and ② a (length-free) description, which often includes ③ the expected skills a candidate should have. As one might imagine, collecting, cleansing, classifying, and then reasoning over these huge amounts of data is a very significant concern for both public and private labor market operators, it should allow for describing trends and dynamics of labor market phenomena from several points of view (e.g., territorial area, emerging occupations, and skills). In such a context, the EU has been making a great effort to define an international skills/occupations classification system (i.e., ESCO[†]) that would represent a lingua franca to labor market analysts and policy makers for studying the labor market dynamics over several countries and overcoming the linguistic boundaries.

In 2015, the CRISP-UNIMIB[‡], in collaboration with the Information Retrieval Laboratory (IR-Lab) within the Department of Computer Science, Systems, and Communications of the University of Milano-Bicocca (UNIMIB), started to work on a European project granted by Cedefop[††], which aims at both building a prototype for analyzing Web job vacancies collected from five EU countries and extracting the requested skills from the data. The rationale behind the project is to turn data extracted from Web job vacancies into knowledge (and thus value) for supporting labor market intelligence. To this end, the well-known knowledge discovery in databases (KDD) process [23] has been applied as a methodological framework. In fact, the project presents some interesting aspects that frame it within the big data panorama—in addition to its relevance for the whole European labor market monitoring system—because it requires dealing with the four "V"s of the big data context: the "volume" of the data (in terms of collected job vacancies growing over time), the "velocity" through which job boards publish new vacancies and close previous ones, the "variety" given by the different data characteristics of each Web source (i.e., semi-structured and unstructured data), and the "veracity," due to the presence of duplicated job vacancies over several sources, or missing information to be identified and addressed. In the following discussion, a process overview is provided, which highlights for each step both the "V"s faced and the technology used.

In the source selection step (quality), 70 Web sources have been ranked according to quality criteria defined by domain experts (e.g., presence of updated posts and territorial granularity). In the data collection step (volume, velocity, variety, and veracity) a modular scraper composed of three distinct components was built, namely ① a downloader for retrieving the Web page, ② an extractor that recognizes the main elements of a job vacancy and stores them in a database, and ③ a monitor that schedules and executes the overall scraping process periodically. This module has been built in house to deal with the high heterogeneity of Web sources using the Spring Framework and Talend for the task orchestration. About 4 million job vacancies have been collected in a trimester over five European countries. The data cleaning and classification task (volume, variety, and veracity) takes care of recognizing duplicate job vacancies and classifying each of them according to the ESCO occupation taxonomy (about 436 occupation items). Notice that the data cleansing is a far-from-straightforward process, as it may affect the believability of subsequent steps (see e.g., Refs. [24–26]). To this end, a machine learning algorithm has been used, as it outperformed other approaches in a domain-dependent benchmark [27] and reached a high level of classification accuracy in the project settings (i.e., from 79% for Germany up to 98% for the Czech Republic).

The classification module was built with custom code using the SciPy framework. The skill extraction task (volume, variety, and veracity) takes charge of identifying and extracting skills from job vacancy descriptions using linguistic models. In this way, the data classified according to the ESCO occupation taxonomy is enriched with information about the skills requested by the employers, thus producing a detailed portrait of the job opportunities advertised through the Web.

Finally, several visualization models were identified using the well-known D3.js visualization library. An example of the end product of this process (focusing only on the Italian labor market data) is WollyBI[‡‡].

In conclusion, this project sheds light on the relevance of applying "intelligent" techniques and data engineering to face the main issues related to big data in a real and domain-specific context. The research findings pave the way for future works in the following directions: first, to automatically group similar occupations on the basis of the skills requested by the employers; second, to represent the collected knowledge via a graph-based model, which is a natural and convenient choice for a large and highly dynamic knowledge base including all the job vacancies (tens of millions of nodes). Then, after the project deployment, a huge amount of data about the Web labor market for some principal European countries is expected to be collected. This would represent a valuable knowledge base that would be beneficial for research activities in the domain of labor market intelligence.

6. Privacy and ethics in big data analytics

The big data originating from the digital breadcrumbs of human activities, sensed as a byproduct of the information and communication technology (ICT) systems that we use every day, record the multiple dimensions of social life: Automated payment systems record the tracks of our purchases; search engines record the logs of our queries on the Web; and wireless networks and mobile devices record the traces of our movements. These big data describing human activities are at the heart of the idea of a "knowledge society," where the understanding of social phenomena is sustained by knowledge extracted from the miners of big data across the various social dimensions, using social mining technologies. Thus, the analysis of our digital traces can create new opportunities to understand complex aspects, such as mobility behaviors, economic and financial crises, the spread of epidemics, and the diffusion of opinions. However, the remarkable opportunities for discovering interesting patterns from these data can be outweighed by the high risk of ethical issues in data processing and analysis and by the ethical consequences of suggestions and predictions. Important ethical risks are: ① privacy violations, when uncontrolled intrusion into the personal data of the subjects occurs, and ② discrimination, when the discovered knowledge is unfairly used in making discriminatory decisions about the (possibly unaware) people who are classified or profiled.

Nevertheless, big data analytics and ethics are not necessarily enemies. In the literature, some works have shown that many practical and impactful services, based on big data analytics, can be designed in such a way that the quality of results can coexist

[†] ESCO is the multilingual classification of European Skills, Competences, Qualifications, and Occupations built on top of the International Standard Classification System (ISCO). ESCO is part of the Europe 2020 strategy.

[‡] The Interuniversity Research Centre on Public Services–University of Milan-Bicocca.

[††] The European Center for the Development of Vocational Training.

[‡‡] http://www.wollybi.com/en/

with the enforcement of ethical requirements. The secret is to develop big data analytics technologies that, by design, enforce ethical value requirements in order to offer safeguards of fairness.

In the context of privacy protection in big data analytics, Monreale et al. [28] proposed the instantiation of the privacy-by-design paradigm, introduced by Ann Cavoukian in the 1990s, to the designing of big data analytical services. This methodology was applied to guarantee privacy in the following fields.

6.1. Privacy in mobility data publishing

Monreale et al. [29] designed a method for the privacy-aware publication of movement data, enabling clustering analysis, which is useful for understanding human mobility behavior in specific urban areas. The released trajectories are made anonymous by a suitable process that realizes a generalized version of the original trajectories. The results obtained with the application of this framework show how trajectories can be anonymized to a high level of protection against re-identification, while preserving the possibility of mining clusters of trajectories, which enables novel and powerful analytic services for info-mobility or location-based services.

6.2. Privacy in data mining outsourcing

Giannotti et al. [30] designed a method for a privacy-aware outsourcing of the pattern mining task. In particular, the results show how a company can outsource the transaction data to a third party and obtain a data mining service in a privacy-preserving manner. In this setting, not only the underlying data but also the mined results (the strategic information) are not intended for sharing and must remain private. The privacy solution proposed in Ref. [27] involves applying an encryption scheme that transforms the original database by the following steps: ① replacing each item by a 1-1 substitution function; and ② adding fake transactions to the database in such a way that each item (itemset) becomes indistinguishable with at least $(k - 1)$ other items (itemsets). On the basis of this simple idea, this framework guarantees that not only individual items, but also any group of items, have the property of being indistinguishable from at least k other groups in the worst case, and actually many more in the average case. This protection implies that the attacker has a very limited probability of guessing the actual items contained either in the transaction data or in the mining results. In contrast, the data owner can efficiently decrypt correct mining results, returned by the third party, with limited computational resources.

6.3. Privacy in distributed analytical systems

Monreale et al. [31] proposed a method for a privacy-aware distributed mobility data analytics, for a situation in which an untrusted central station collects some aggregate statistics computed by each individual node that observes a stream of mobility data. The central station stores the received statistical information and computes a summary of the traffic conditions of the whole territory, based on the information collected from data collectors. The proposed framework guarantees privacy protection at the individual level by applying a well-known privacy model called "differential privacy." In particular, the privacy technique perturbs nodes' mobility data before transmitting them to the untrusted central station.

6.4. Discrimination discovery from data and discrimination prevention

In the context of discrimination data analysis, two main lines of research are being pursued (see Ref. [32] for a survey). Discrimination discovery from data consists in the actual discovery of discriminatory situations and practices hidden in a large amount of historical decision records. A process for direct and indirect discrimination discovery on social groups using classification rule mining and filtering was originally proposed. The process is guided by legally grounded measures of discrimination, possibly including statistical tests of confidence [32]. Individual discrimination has instead been modeled with a k-NN approach, and applied to a real case study in research project funding [33].

Discrimination prevention consists of removing bias from training data and from learning algorithms that may lead to predictive models that may make (possibly autonomous) discriminatory decisions. Data sanitization for discrimination prevention has been investigated in Ref. [34], by first reducing the t-closeness model of privacy to a model for non-discrimination, and then by adapting state-of-the-art data sanitization methods for t-closeness. An approach dealing with both privacy and discrimination sanitization is in Ref. [35]. Regarding learning algorithms, a modified voting mechanism of rule-based classifiers to reduce the weight of possibly discriminatory rules has been proposed [32].

7. Making big data usable

7.1. Entity resolution for big data

The Web has become a valuable source of structured and semi-structured data. A huge amount of high-quality relational data can be extracted from HTML tables [36], and, with the advent of the Web of data, the amount of semi-structured data publicly available as linked data is exponentially growing [37]. These data are characterized by high volume and variety and rapid changes, and both their veracity and quality are often an issue [38,39]. For these reasons, this kind of data is commonly identified as "big data." The true potential of these data is expressed when different sources are integrated, as demonstrated by recent efforts in mining the Web to extract entities, relationships, and ontologies to build large-scale general-purpose knowledge bases, such as Freebase and Yago [40]. For enterprises, government agencies, and researchers in large scientific projects, these data can be even more valuable if integrated with the data that they already own and that are typically subject to traditional data integration processes.

Being able to identify records that refer to the same entity is a fundamental step to make sense of these data. Generally, in order to perform entity resolution (ER), traditional techniques require a schema alignment between data sources. Unfortunately, big data is typically characterized by high heterogeneity, noise, and very large volume, making traditional schema alignment techniques no longer applicable. For example, Google Base contains over 10 000 entity types that are described with 100 000 unique schemata; in such a scenario, performing and maintaining a schema alignment is impracticable [41].

Recently, two kinds of techniques have been proposed to address these issues: ① techniques that renounce exploiting schema information and rely exclusively on redundancy to limit the chance of missing matches [42–44]; and ② techniques that extract vague schema information directly from the data, which are useful for ER, without performing a traditional schema alignment [45]. The latter results are the most promising, yet least explored. In fact, following their suggested direction, it will be possible to support schema-based ER techniques for big data, which guarantee high recall and precision, without performing the unbearable traditional schema alignment step.

7.2. Big data exploration

In the big data era, new user interfaces are required to interact with the huge amount of data we are able to collect; otherwise, the users will be overwhelmed by data. In Ref. [46], a solution is presented, which helps users to focus their attention on a small set of relevant data, inferred from the user's selection using a Bayesian approach. In our experiments, we examined a method to infer relevant information utilizing the user input in a big data context.

Faceted browsing [47], enhanced with Bayesian networks, is used as probabilistic models to infer valuable information for the user, through the analysis of their selection. Faceted browsing is a technique for performing data exploration by applying dynamic filters in multiple steps: Each time a filter is applied, the results are displayed to the user, who can apply additional filters or modify existing ones. At each step, the displayed filters and the values inside the filters might be different.

The proposed method is effective for exploring data in a big data context, where the number of attributes and their values are huge. In other words, the advantage offered by faceted browsing is the dynamicity of filters. Moreover, in order to dynamically obtain the most valuable filters for the user, it is necessary to infer them by utilizing the user's current selections. Thus, by means of the analysis of a user's selection with a graphical Bayesian network probabilistic model, it is possible to infer the most valuable filters for him/her. Graphical models are preferred, since they are easy to understand, verify, and interpret the results. In this context, the variables inside a Bayesian network are the attributes of a dataset. Bayesian networks are exploited to infer the relationships among these attributes, to calculate the probability of correlation between a user's selection and the other attributes of the network, and then to display the most relevant attributes as filters. In addition, it is possible to infer similar or dissimilar values in the filters, in order to avoid displaying too many values. To conclude the process, only the top five similar and dissimilar values are shown to the user.

8. Conclusions

This article has presented some of the many academic research activities on big data performed in Italy, by covering both applications aiming to improve various aspects of people's lives and two general, important issues of privacy and big data usability. It shows a prolific academic research community, which is ready to face all challenges currently posed by the volume, velocity, variety and veracity of big data. The next stage is a tighter cooperation with industry in order to face together the most relevant challenge: creating value from big data. In this sense, the participation to the implementation of the European strategic research agenda defined by the Big Data Value Association, also through the support of the CINI "Big Data" Laboratory, will be crucial.

Compliance with ethics guidelines

Sonia Bergamaschi, Emanuele Carlini, Michelangelo Ceci, Barbara Furletti, Fosca Giannotti, Donato Malerba, Mario Mezzanzanica, Anna Monreale, Gabriella Pasi, Dino Pedreschi, Raffele Perego, and Salvatore Ruggieri declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Europe Big Data market 2015–2020 [Internet]. New York: PR Newswire Association LLC.; c2016 [updated 2016 May 30, cited 2016 Jun 12]. Available from: http://www.prnewswire.com/news-releases/europe-big-data-market-2015---2020-300276656.html.

[2] Furletti B, Gabrielli L, Renso C, Rinzivillo S. Analysis of GSM calls data for understanding user mobility behavior. In: Hu X, Lin TY, Raghavan V, Wah B, Baeza-Yates R, Fox G, et al., editors Proceedings of the 2013 IEEE International Conference on Big Data; 2013 Oct 6–9; Santa Clara, CA, USA; 2013. p. 550–5.

[3] Furletti B, Gabrielli L, Renso C, Rinzivillo S. Pisa tourism fluxes observatory: deriving mobility indicators from GSM call habits. In: Proceedings of the 3rd International Conference on the Analysis of Mobile Phone Datasets; 2013 May 1–3; Cambridge, MA, USA; 2013.

[4] Gabrielli L, Furletti B, Trasarti R, Giannotti F, Pedreschi D. City users' classification with mobile phone data. In: Ho H, Ooi BC, Zaki MJ, Hu X, Haas L, Kumar V, et al., editors Proceedings of the 2015 IEEE International Conference on Big Data; 2015 Oct 29–Nov 1; Santa Clara, CA, USA; 2015. p. 1007–12.

[5] Furletti B, Gabrielli L, Giannotti F, Milli L, Nanni M, Pedreschi D. Use of mobile phone data to estimate mobility flows. Measuring urban population and inter-city mobility using big data in an integrated approach. In: Proceedings of the 47th SIS Scientific Meeting of the Italian Statistical Society; 2014 Jun 11–13; Cagliari, Italy; 2014.

[6] Nanni M, Trasarti R, Furletti B, Gabrielli L,Van Der Mede P, De Bruijn J, et al. Transportation planning based on GSM traces: a case study on ivory coast. In: Nin J, Villatoro D, editors Citizen in sensor networks. Cham: Springer International Publishing; 2014. p. 15–25.

[7] Pappalardo L, Simini F, Rinzivillo S, Pedreschi D, Giannotti F, Barabási AL. Returners and explorers dichotomy in human mobility. Nat Commun 2015;6:8166.

[8] Pappalardo L, Pedreschi D, Smoreda Z, Giannotti F. Using big data to study the link between human mobility and socio-economic development. In: Ho H, Ooi BC, Zaki MJ, Hu X, Haas L, Kumar V, et al., editors Proceedings of the 2015 IEEE International Conference on Big Data; 2015 Oct 29–Nov 1; Santa Clara, CA, USA; 2015. p. 871–8.

[9] Wang D, Pedreschi D, Song C, Giannotti F, Barabási AL. Human mobility, social ties, and link prediction. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2011 Aug 21–24; San Diego, CA, USA; 2011. p. 1100–8.

[10] Trasarti R, Olteanu-Raimond AM., Nanni M, Couronné T, Furletti B, Giannotti F, et al. Discovering urban and country dynamics from mobile phone data with spatial correlation patterns. Telecommu Policy 2015;39(3–4):347–62.

[11] Liu W, Park EK. Big data as an e-health service. In: Proceedings of the 2014 IEEE International Conference on Computing, Networking and Communications; 2014 Feb 3–6; Honolulu, HI, USA; 2014. p. 982–8.

[12] Gini R, Francesconi P, Mazzaglia G, Cricelli I, Pasqua A, Gallina P, et al. Chronic disease prevalence from Italian administrative databases in the VALORE project: a validation through comparison of population estimates with general practice databases and national survey. BMC Public Health 2013;13(1):15.

[13] Directive 2009/28/EC of the European Parliament and of the Council on the promotion of the use of energy from renewable sources and amending and subsequently repealing Directives 2001/77/EC and 2003/30/EC. Official Journal of the European Union L 140; 2009 Jun 5. p. 16–47.

[14] Ioakimidis CS, Oliveira LJ, Genikomsakis KN. Wind power forecasting in a residential location as part of the energy box management decision tool. IEEE Trans Ind Inform 2014;10(4):2103–11.

[15] Bessa RJ, Miranda V, Gama J. Entropy and correntropy against minimum square error in offline and online three-day ahead wind power forecasting. IEEE Trans Power Syst 2009;24(4):1657–66.

[16] Ceci M, Cassavia N, Corizzo R, Dicosta P, Malerba D, Maria G, et al. Innovative power operating center management exploiting big data techniques. In: Proceedings of the 18th International Database Engineering & Applications Symposium; 2014 Jul 7–9; Porto, Portugal. New York: ACM; 2014. p. 326–9.

[17] Ceci M, Corizzo R, Fumarola F, Ianni M, Malerba D, Maria G, et al. Big data techniques for supporting accurate predictions of energy production from renewable sources. In: Proceedings of the 19th International Database Engineering and Applications Symposium; 2015 Jul 13–15; Yokohama, Japan New York: ACM; 2015. p. 62–71.

[18] Bofinger S, Heilscher G. Solar electricity forecast—approaches and first results. In: Proceedings of the 21st European Photovoltaic Solar Energy Conference; 2006 Sep 4–8; Dresden, Germany; 2006. p. 4–8.

[19] Pelland S, Galanis G, Kallos G. Solar and photovoltaic forecasting through post-processing of the Global Environmental Multiscale numerical weather prediction model. Prog Photovoltaics 2013;21(3):284–96.

[20] Sharma N, Sharma P, Irwin DE, Shenoy PJ. Predicting solar generation from weather forecasts using machine learning. In: Proceedings of the 2011 IEEE International Conference on Smart Grid Communications; 2011 Oct 17–20; Brussels, Belgium; 2011. p. 528–33.

[21] Stojanova D, Ceci M, Appice A, Džeroski S. Network regression with predictive clustering trees. Data Min Knowl Disc 2012;25(2):378–413.

[22] Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I. Spark: cluster computing with working sets. In: Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing; 2010 Jun 22–25; Boston, MA, USA. Berkeley: USENIX Association; 2010. p. 1765−73.

[23] Fayyad U, Piatetsky-Shapiro G, Smyth P. The KDD process for extracting useful knowledge from volumes of data. Commun ACM 1996;39(11):27–34.

[24] Boselli R, Cesarini M, Mercorio F, Mezzanzanica M. Planning meets data cleansing. In: Proceedings of the 24th International Conference on Automated Planning and Scheduling; 2014 Jun 21–26; Portsmouth, NH, USA; 2014. p. 439−43.

[25] Mezzanzanica M, Boselli R, Cesarini M, Mercorio F. Data quality sensitivity analysis on aggregate indicators. In: Helfert M, Francalanci C, Felipe J, editors Proceedings of the International Conference on Data Technologies and Applications; 2012 Jul 25−27; Rome, Italy; 2012. p. 97−108.

[26] Mezzanzanica M, Boselli R, Cesarini M, Mercorio F. A model-based evaluation of data quality activities in KDD. Inform Process Manag 2015;51(2):144−66.

[27] Amato F, Boselli R, Cesarini M, Mercorio F, Mezzanzanica M, Moscato V, et al. Challenge: processing web texts for classifying job offers. In: Kankanhalli MS, Li T, Wang W, editors Proceedings of the 2015 IEEE International Conference on Semantic Computing; 2015 Feb 7−9; Anaheim, CA, USA; 2015. p. 460−3.

[28] Monreale A, Rinzivillo S, Pratesi F, Giannotti F, Pedreschi D. Privacy-by-design in big data analytics and social mining. EPJ Data Sci 2014;3(1):10.

[29] Monreale A, Andrienko G, Andrienko NV, Giannotti F, Pedreschi D, Rinzivillo S, et al. Movement data anonymity through generalization. Trans Data Privacy 2010;3(2):91−121.

[30] Giannotti F, Lakshmanan LVS, Monreale A, Pedreschi D, Wang H. Privacy-preserving mining of association rules from outsourced transaction databases. IEEE Syst J 2013;7(3):385–95.

[31] Monreale A, Wang WH, Pratesi F, Rinzivillo S, Pedreschi D, Andrienko G, et al. Privacy-preserving distributed movement data aggregation. In: Vandenbroucke D, Bucher B, Crompvoets J, editors Geographic information science at the heart of Europe. Cham: Springer International Publishing; 2013. p. 225−45.

[32] Romei A, Ruggieri S. A multidisciplinary survey on discrimination analysis. Knowl Eng Rev 2014;29(5):582–638.

[33] Romei A, Ruggieri S, Turini F. Discrimination discovery in scientific project evaluation: a case study. Expert Syst Appl 2013;40(15):6064–79.

[34] Ruggieri S. Using *t*-closeness anonymity to control for non-discrimination. Trans Data Privacy 2014;7(2):99–129.

[35] Hajian S, Domingo-Ferrer J, Monreale A, Pedreschi D, Giannotti F. Discrimination- and privacy-aware patterns. Data Min Knowl Disc 2015;29(6):1733–82.

[36] Cafarella MJ, Halevy A, Wang ZD, Wu E, Zhang Y. WebTables: exploring the power of tables on the web. In: Proceedings of the Very Large Database Endowment; 2008 Aug 23–28; Auckland, New Zealand; 2008. p. 538–49.

[37] Bizer C, Heath T, Berners-Lee T. Linked data: the story so far. In: Sheth A, editor Semantic services, interoperability and web applications: emerging concepts. Hershey: IGI Global; 2011. p. 205–27.

[38] Batini C, Rula A, Scannapieco M, Viscusi G. From data quality to big data quality. J Database Manage 2015;26(1):60–82.

[39] Firmani D, Mecella M, Scannapieco M, Batini C. On the meaningfulness of "Big Data Quality". Data Sci Eng 2016;1(1):6−20.

[40] Dong XL, Srivastava D. Big data integration. In: Proceedings of the Very Large Databases Endowment; 2013 Aug 26−30; Trento, Italy; 2013. p. 1188−9.

[41] Madhavan J, Jeffery SR, Cohen S, Dong XL, Ko D, Yu C, et al. Web-scale data integration: you can afford to Pay As You Go. In: Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research; 2007 Jan 7–10; Asilomar, CA, USA; 2007. p. 342–50.

[42] Papadakis G, Ioannou E, Palpanas T, Niederée C, Nejdl W. A blocking framework for entity resolution in highly heterogeneous information spaces. IEEE Trans Knowl Data En 2013;25(12):2665–82.

[43] Papadakis G, Koutrika G, Palpanas T, Nejdl W. Meta-blocking: taking entity resolution to the next level. IEEE Trans Knowl Data En 2014;26(8):1946–60.

[44] Papadakis G, Papastefanatos G, Koutrika G. Supervised meta-blocking. In: Proceedings of the Very Large Databases Endowment; 2014 Sep1–5; Hangzhou, China.; 2014. p. 1929–40.

[45] Bergamaschi S, Ferrari D, Guerra F, Simonini G. Discovering the topics of a data source: a statistical approach. In: Proceedings of the Workshop on Surfacing the Deep and the Social Web Co-located with the 13th International Semantic Web Conference; 2014 Oct 19; Trentino, Italy; 2014.

[46] Bergamaschi S, Simonini G, Zhu S. Enhancing big data exploration with faceted browsing. In: Proceedings of the 10th Scientific Meeting of Classification and Data Analysis Group; 2015 Oct 8-10; Cagliari, Italy; 2015.

[47] Fagan JC. Usability studies of faceted browsing: a literature review. Inform Technol Libr 2010;29(2):58–66.