How data mining and machine learning evolved from Relational Data Base to Data Science

G. Amato[1], L. Candela, D. Castelli[1], A. Esuli[1], F. Falchi[1], C. Gennaro[1], F. Giannotti[1], A. Monreale[2], M. Nanni[1], P. Pagano, L. Pappalardo[2], D. Pedreschi[2], F. Rabitti[1], S. Rinzivillo[1], F. Sebastiani[1], G. Rossetti[2], S. Ruggieri[2], and M. Tesconi[3]

[1] ISTI - CNR, Pisa, Italy
[2] University of Pisa
[3] IIT-CNR, Pisa, Italy

1 Introduction

During the last 35 years, data management principles such as physical and logical independence, declarative querying and cost-based optimization have led to profound pervasiveness of relational databases in any kind of organization. More importantly, these technical advances have enabled the first round of business intelligence applications and laid the foundation for managing and analyzing Big Data today. The 90's have been exceptional years for the invention and development of solid data mining and machine learning algorithms [1,2,65] building on existing statistical and artificial intelligence theories. Open and proprietary software libraries and analytical platforms have bloomed in parallel with the development of a robust methodological approach to the development of analytical processes capable of extracting valuable knowledge out of large masses of data: the Knowledge Discovery in Databases (KDD) [39]. When, the data deluge began, the KDD technologies were well prepared so that the new advances stimulated by the many novel challenges and opportunities associated with Big Data took place in parallel very effective field demonstrations in a wide array of domains, thus activating a virtuous cycle between innovation and research.

The data deluge has been really impressive: digital technology has become ubiquitous and very much part of public and private organizations and individuals. People and things have become increasingly interconnected. Smartphones, buildings, cities, vehicles and other environments and devices have been filled with digital sensors, all of them creating evermore data. New high-throughput scientific instruments, telescopes, satellites, accelerators, supercomputers, sensor networks, and running simulations have generated and are generating massive amounts of data.

Big data have been blossoming together with the hope to harness the knowledge they hide to solve the key problems of society, business and science. However, turning an ocean of messy data into knowledge and wisdom is an extremely challenging task. Heterogeneity, scale, timeliness, complexity, and privacy problems with Big Data are the key issues to be addressed at all phases of the pipeline that can create value from data.

Big Data is not natively in structured format; for example, tweets and blogs are weakly structured pieces of text, while images and video are structured for storage and display, but not according its semantic content to enable search: transforming such content into a structured format for later analysis has been and is a major challenge. The value of data explodes when it can be linked with other data, thus data integration is a major creator of value. Since most data is directly generated in digital format today, we have the opportunity and the challenge both to influence the creation to facilitate later linkage and to automatically link previously created data.

In this context, at the end of the 90s a new analytical trend joined data mining and machine learning: the emergence of *network science* [16]. Once again, the availability of large graph data emerging from the web has allowed to discover general patterns and statistical laws regulating statics and dynamics of complex networks.

Another relevant impact of big data is the opportunity to observe and measure how our society intimately works: the digital breadcrumbs of human activities carried the capacity to scrutinize the ground truth of individual and collective behaviour at an unprecedented detail. Multiple dimensions of our social life have been increasingly "proxied" by big data: automated payment systems record the tracks of our purchases; search engines record the logs of our queries on the web; wireless networks and mobile devices record the traces of our movements; social media record the traces of our opinions and emotions; social networks record the traces of our interactions. This new scenario took the name of *social mining* and it is clear that such challenge requires high-level analytics, modeling and reasoning across all the social dimensions above.

This chapter proposes an account of the scientific and technical evolution of data mining and machine learning from relational data bases to data science, focusing on "making sense" of data generated as by-product of ICT mediated human activities, i.e. on the analytical methods and process, intentionally neglecting the amazing advances on the efficiency and scalability of the algorithms as well as their ability to deal with massive streaming data. The chapter tells the story of this evolution through the research achievements of a network of research labs in Pisa across the CNR and the University, which contribute to the birth and life of the Italian database community of SEBD. The next sections discuss are dedicated to the main trends and results of these groups in the following areas. **Mobility data analysis** leverages the spatio-temporal dimensions of big data to the purpose of understanding human mobility behavior, evolutionary patterns, daily activity patterns, geographic patterns. **Social network analysis** studies the architecture of interpersonal relationships, with the purpose of understanding the structure and the dynamics of the fabric of human society. **Multimedia media mining** methods for making sense of heterogeneous data, sensed from different on line sources: tweets, mails, blogs, web pages, link structures, videos etc., to the purpose of extracting the hidden semantics from them. **Sentiment mining**. At the crossroads of natural language processing and information retrieval, a key topic is opinion/sentiment mining, aimed at harnessing

the emotional content of user generated texts. The last two sections present two activities that have a transversal impact. First, privacy aware analytics to prevent "by-design" the risks of invading the sphere of personal information and the reflection on the ethical consequences of predictive analytics. Second, the new opportunities that data infrastructures and virtual research environments bring to data scientists.

2 Mobility data analysis

In the last decades, the wide-spread availability of geo-localization devices and the technologies to store and analyze the data they generate had a huge impact on the research areas dealing with spatial and spatio-temporal data. GPS traces and other forms of mobility data quickly became a focus for researchers from several disciplines, especially in the domain of data mining. These mobility data provide a new powerful social microscope, which may help us understand human mobility, and discover the hidden patterns and models that characterize the trajectories humans follow during their daily activity.

Revisitations of standard data mining problems and techniques quickly appeared, many of them coming from the important contributions provided by the Italian community in the last 15 years: trajectory patterns, i.e. frequent patterns describing sequences of places and possibly timings that are common to a large number of input trajectories; trajectory clustering, i.e. grouping similar trajectories [10] or trajectory segments [53] into homogeneous groups; location prediction, i.e. forecasting the position that a moving object will have in the near future [57]; recognition of movement activity, i.e. associating to a trajectory or to a stop area the activity it was aimed to [66], such as going to work, leisure, shopping, etc.

In the rest of this section we will mention two examples that had a great impact on the research community, in terms of references and applications that stemmed from them: Trajectory Patterns and Mobility Profiles.

2.1 Trajectory patterns

In some contexts, the moving objects we are examining might act in a similar way, even if they are not spatially located together. For instance, similar daily routines might lead several individuals to drive their car along the same routes, even if they leave home at very different hours of the day. Or, tourists that visit a city on different days of the year might actually visit it in the same way – for instance by visiting the same places in the same order and spending there approximately the same amount of time – because they simply share interests and attitude. The kind of questions we might try to answer in this cases is: *Are there groups of objects that perform a sequence of movements, with similar timings though possibly during completely different moments?* Accordingly, ***T-Patterns*** [47] (abbreviation of *Trajectory patterns*) provide sequences of spatial locations with typical transition times, such

as Railway Station $\stackrel{15min}{\longrightarrow}$ Museum $\stackrel{2h15min}{\longrightarrow}$ Castle Square. This might represent the typical behavior of tourists that rapidly reach a museum from the railway station and spend there about two hours before getting to the adjacent square.

The set of spatial regions to be used to form patterns is a major parameter of the method, i.e., the spatial extension of "Railway Station" and any other place considered relevant for the analysis. However, the algorithm proposed also contains heuristics to automatically define such regions, based on coverage of dense areas, in case there is no domain expert to provide them.

T-Patterns represented the first attempt to automatically infer from raw GPS traces an higher abstraction of movement, capturing key places (the regions in the patterns) and temporal evolution. This information has been later exploited as building block of various applications, such as prediction (the WhereNext method [57], which predicts the next most likely area a moving object will visit), or the identification of hot routes in the city [46].

2.2 Mobility profiles

Despite the great attention that the analysis of individual trajectories attracted, for a very long time the individuals themselves have not been considered as a relevant subject of analysis. **Mobility profiles** [79] represent the first clear step on the opposite direction, by analysing individuals (rather than just large groups) with the purpose of understanding systematic mobility, as opposed to occasional movements, which is fundamental in some mobility planning applications, e.g. public transport.

The objective is to use the set of trips of an individual user to find his/her routine behaviors. That is realized by grouping together similar trips based on concepts of spatial distance and temporal alignment, with corresponding thresholds for both the spatial and temporal components of the trips. In order to be defined as *routine*, a behavior needs to be supported by a significant number of similar trips of the user. The technology adopted to achieve that is a clustering algorithm that groups together the similar trajectories, each cluster representing a routine [10]. In particular, the algorithm combines density-based methods (known to deal well with non-spherical clusters and noisy data, both being typical features of trajectory data) with bisecting k-means (used to obtain compact clusters). Each group obtained is then summarized by its central element.

Individual Mobility profiles enable several applications, ranging from deeper traffic analyses (indeed, the traffic traversing a given area can now be described also by a systematicity index, measuring the percentage of trips that are routinaries for the individuals involved) and the creation of predictive models (as in the case of MyWay [78], where ongoing trips are compared against the user's routines, and in case of match they are used to predict how the trip will continue) to services like carpooling [48].

2.3 The borders of human mobility.

The problem of discovering the geographic borders from human mobility at the low spatial resolution of municipalities or counties is a far reaching problem, motivated by providing policy makers with suggestions about the best administrative partitions for the government of the territory. In [26,67], we adopt a social network analysis view to mobility data to reach a better understanding of human mobility patterns, leveraging the underlying, hidden connections that human mobility establishes among different places. Starting from a given zoning of the territory, tessellated into census zones, we construct a network whose nodes are the zones and the weighted edges between any two zones represent the number of travels originating in the first and ending in the second. The analysis phase consists in discovering densely connected sub-graphs in this network by means of a community detection method, thus highlighting groups of zones that are highly connected by many travels compared to the lower connectivity among different modules. This an example how a network mining method can be adopted to reveal the hierarchical structure of a complex phenomenon, and highlight the thresholds at which we separate *macro-, meso- and micro-levels* of the system.

2.4 Returners and Explorers

Another interesting line of research has been at the crossroad of mobility data mining and network science. Network science is aimed at discovering the global models of complex social phenomena, by means of statistical macro-laws governing basic quantities, which show the behavioral diversity in society at large. Data mining is aimed at discovering local patterns of complex social phenomena, by means of micro-laws governing behavioral similarity or regularities in sub-populations. The objective of combining micro and macro laws has has been pursued in [62] where taking advantage of massive digital traces of human whereabouts a series of novel insights on the quantitative patterns characterizing human mobility have been discovered and used to anchor to reality the abstract models of human mobility. Our work starts from the recent consensus on the fact that the considerable variability in the characteristic travelled distance of individuals coexists with a high degree of predictability of their future locations. Here we shed light on this surprising coexistence by systematically investigating the impact of recurrent mobility on the characteristic distance travelled by individuals. Using both mobile phone and GPS data, we discover the existence of two distinct classes of individuals: returners and explorers. As existing models of human mobility cannot explain the existence of these two classes, we develop more realistic models able to capture the empirical findings. Finally, we show that returners and explorers play a distinct quantifiable role in spreading phenomena and that a correlation exists between their mobility patterns and social interactions.

2.5 Sociometer: classification of city users and flows

One very promising source of mobility information are mobile phones traces, most commonly collected in the form of Call Detail Records (CDRs), i.e. records of the phone calls performed that describe the starting time and location (in terms of antenna connected in that moment) of the call. In this context, an analysis method named *Sociometer* was developed, aimed to associate to each user her role w.r.t. a specific area, such as *resident*, *visitor*, etc.

Adopting a vision similar to the mobility profiles described above, our work started from the analysis of the single users' behavior. The approach summarizes the CDR data of each user through a temporal distribution of her calls within the spatial area under consideration, measuring the percentage of days that her was seen at different hours of the day (grouped into three intervals of around 8 hours each), in different days of the week (grouped into *week-days* and *week-ends*), in different weeks. The basic idea is that different city users will produce different kinds of temporal distribution, for instance residents will most likely be present on all the time slots, while commuters will be seen only during working hours/days.

The personal *fingerprints* are clustered to identify the most relevant calling patterns, which are then classified through a standard K-NN classification schema. Earlier versions of the solution were based on a manual labelling of the relevant calling patterns found [41], while most recent ones compare them against a pre-defined set of representative distributions, called *archetypes* [42].

3 Social Network Analysis

Nowadays Complex Networks are pervasively used to model and describe the behaviors of a wide range of real world phenomena. Social relationships, biological interactions, transportation, commercial exchanges are only few of the several scenarios usually studied with the support of instruments borrowed by graph theory. Countless problems are formulated, or can be formulated, upon such structures: Community Discovery, Link Prediction, Tie Strength estimation are only few of them. Among all the fields that emerged in the last decades Social Network Analysis, SNA, is the one that makes use of graph mining techniques to understand human behaviors. SNA research has certainly be facilitated by the ever-growing popularity of online social network platforms data available. Such unprecedented sources of human generated data naturally modelled by the tools and theories offered by graph theory have lead to the rising of this novel field of research. Among the vast SNA literature, our research group has effectively contributed to the following themes: Multidimensional network analysis, Community Discovery, Network Analytics & Mobility and analysis of diffusive patterns.

3.1 Multidimensional Networks

Most real life networks are intrinsically multidimensional, and some of their properties may be lost if the different dimensions are not taken into account. In

other cases, it is natural to derive multiple dimensions connecting a set of nodes from the available data to the end of analyzing some phenomena. In order to study this complex scenario a framework that extends the classical graph theory is needed. Reasoning on multidimensional networks seems clear that the usual graph model is not enough to represent all the available information. In our work *"Foundations of Multidimensional Network Analysis"* [17], using a multigraph representation, we proposed and evaluated on real datasets a multidimensional framework able to capture the interplay among dimensions and to overcome some limits that made the classical monodimensional measures unsuitable in this complex scenario. Such framework was then extended, in [61] where we formulate an approach to estimate tie strength on multidimensional networks and validate it on a multigraph built upon the social relationships of users interacting on three different online platform, namely Facebook, Twitter and Foursquare. Moreover, in [69] we proposed and evaluated a set of Link Prediction approach specifically tailored for multidimensional networks.

3.2 Community Discovery

The problem of identifying communities in complex networks is very popular among network scientists, as witnessed by an impressive number of valid works in this field. Traditionally, a community is defined as a dense subgraph, in which the number of edges among the members of the community is significantly higher than the outgoing edges. Our survey [25] explores all the most popular techniques to find communities in complex networks and categorize them into eight main categories: Feature Distance, Internal Density, Bridge Detection, Closeness, Structure Definition, Link Clustering, Meta Clustering and Diffusion. In [27] we propose a bottom-up approach to efficiently extract overlapping communities: DEMON. DEMON leverages the nodes perspective to identify meaningful network substructures: it works by identify local-communities at the ego-network level exploiting label propagation and then merging them in an incremental fashion. Our approach has been used as a proxy for users homophily to support network quantification tasks [55]; as filter to reduce the computational cost of Link Prediction approaches [70]; as well as to bound set of Skype users while searching a network driven methodology to relate service usage to network position [71]. Moreover, in order to cope with the evolving nature of interaction networks, we proposed an online dynamic community discovery algorithm, TILES [72], able to track community life cycles as new perturbations appears in the network (i.e. appearance/ vanishing of nodes as well as edges).

4 Sentiment analysis

A large proportion of the data that is generated daily, and that needs to be processed by search and mining algorithms, is of a textual, non-structured nature; these data have traditionally been the domain of information retrieval and

text mining. After the advent of the so-called "Web 2.0", a lot of textual content is user-generated, and its nature is not purely descriptive: that is, it is not confined to describing facts or states of affairs in an objective, detached way, but is instead rich in subjective, opinionated content. Harnessing the opinions and emotions expressed by the authors of these textual contents is the object of *sentiment analysis* (also known as *opinion mining*), an area at the crossroads of natural language processing and information retrieval that has blossomed in the mid years of the past decade, and that has been receiving increased attention, from industry and the scientific community alike, ever since.

4.1 Automatically expanding sentiment lexicons

Possibly the most important task underlying attempts to tap into this kind of data is *sentiment classification*, the task of classifying an item of user-generated content (UGC – e.g., a tweet, a product review, a post on a social networking service) according to the sentiment it conveys (or opinion it expresses) about a certain entity. While this shares many characteristics with the task of classifying text by topic, the traditional "bag of words" (BoW) approach to representing textual content cannot be used for classifying text by sentiment: to see why, simply consider the fact that two sentences such as "A horrible hotel in a beautiful town" and "A beautiful hotel in a horrible town" would be assigned the same class if relying on a BoW representation, while they convey radically different sentiment. As a result, classification by sentiment fundamentally relies on the availability of a *sentiment dictionary*, i.e., an online dictionary where lexical entries (e.g., words, or word senses) are tagged in terms of whether they convey a sense of positivity (e.g., "truthful", "sublime") or negativity (e.g., "inaccurate", "pathetic"). However, manually curated sentiment dictionaries characterised by a high coverage of the language rarely exist in practice, especially for less resourced languages. As a result, our group investigated a number of language-independent methods for automatically tagging by sentiment existing online dictionaries.

A first method we developed was based on gloss classification, i.e., on classifying a lexical entry as positive or negative by classifying the textual definition of the entry ("gloss"); the method was first applied to classifying words according to the positive vs. negative dichotomy [38], and later extended to also identify neutral words (i.e., words that convey no sentiment; e.g., "inorganic", "quadratic") [35]. This method was deployed in practice in order to tag the English-language version of WordNet; the result was SentiWordNet [37], a sentiment lexicon now routinely used by hundreds of research groups worldwide.

A second method we later developed was based on random walks, and assumed that a positive (resp., negative) word being defined (the *definiendum*) is defined by mostly using positive (resp., negative) words in the gloss (the *definiens*). By assuming that positivity and negativity "flow" along the links connecting the definiendum with the words contained in the definiens, random walks on the word graph can be used for performing fine-grained computations of how positive/negative a word in a dictionary is. This method was applied

to refining SentiWordNet; this led to a more accurately tagged version, called SentiWordNet 3.0 [15], which is the version now currently available.

4.2 Cross-Lingual and Cross-Domain Sentiment Classification

Cross-lingual sentiment classification is the task of classifying by sentiment text expressed in a target language (e.g., Urdu) when training data are available only for a source language (e.g., English). *Cross-domain* sentiment classification instead refers to sentiment classification of texts about a target domain (e.g., reviews of books) when training data are available only for a source domain (e.g., reviews about CDs). In [60] we have developed a technique that can tackle cross-language *and* cross-domain sentiment classification at the same time (e.g., classifying reviews of books in Urdu when only training reviews of CDs in English are available). The technique, called *Distributional Correspondence Indexing* (DCI), leverages the "distributional hypothesis", i.e., the hypothesis that words with similar meanings tend to occur in the same contexts. DCI derives term representations in a vector space common to both languages/domains where each dimension reflects its distributional correspondence to a pivot, i.e., to a highly predictive term that behaves similarly across languages/domains. Experiments show that DCI obtains better performance than current state-of-the-art techniques for cross-lingual and/or cross-domain sentiment classification.

4.3 Sentiment Quantification

While sentiment classification is important, in [36] we argued that, in many cases of applicative interest (e.g., when analysing tweets or product reviews), the final goal is often not the classification of individual items, but the estimation of the percentage of items that belong to a certain class; in other words, in these cases we are interested not in sentiment classification, but in sentiment *quantification*.

Research has shown that quantification is best tackled by quantification-specific algorithms, and not by using standard classification algorithms followed by counting the number of items that have been assigned the class. In [44,43] we conducted an extensive analysis of existing quantification algorithms as applied to analysing tweets by sentiment; the results confirmed that applying standard classification technology when quantification is the real goal, is suboptimal. Similar conclusions were reached when, instead of standard multi-class quantification, we tackled *ordinal* quantification [30], i.e., the task characterized by a set of classes on which a total order is defined.

5 Multimedia Analysis

Content-based Multimedia Information Retrieval (CBMIR) on a very large scale has been a very active multidisciplinary research field during the last 25 years. Multimedia retrieval involves topics ranging from similarity search, metric access methods and big data, to features extraction, deep learning and smart cameras.

The explosion of multimedia data caused by the diffusion of mobile devices and social media, has increased the relevance of this topic for both industries and governments. We show that the combination of state-of-the-art data structures and deep neural networks allows multimedia analysis that have been considered unachievable for many years because of issues such as semantic gap and curse of dimensionality.

In 2016, a benchmark consisting of 97M deep features[4] extracted from the Yahoo Creative Commons 100M (YFCC100M) dataset [77] was presented and two approximate similarity search techniques were tested on it [8]. In this Section we start from this recent result, to discuss the most relevant research results of the last 25th years that made this possible.

Starting from the CoPhIR dataset dating back to 2009, large datasets have been created using the multimedia shared by users on social media [18]. The proliferation of easily and quickly accessible social media data can be used by researchers for many different purposes. For example, such data has already proved useful for many scenarios such as that of emergency management [13] [12], intelligence [3], eHealth [31], and social networks security [28] [29], to name but a few. In recent years, we have observed the explosion of image-sharing services such as Flickr and Instagram. For instance, Instagram has 600 Million Monthly Users and it was estimated that about 85 million photos are shared everyday Since by sharing photos, users could also express opinions or sentiments, social media images provide a potentially rich source for understanding public opinions.

The features extracted from the images in [8] are the activations of an hidden layer of a Convolutional Neural Networks. This information, automatically extracted from pre-trained deep neural networks, has recently show outstanding results [50], rapidly becoming state of the art in many computer vision applications that have used global (e.g., MPEG-7 Visual Descriptors) and local features (e.g., SIFT, SURF, BRIEF) for decades. Moreover, deep learning [52] is allowing tasks that were not even considered before. In [21], as an example, the authors presented a deep learning based method for searching in a visual feature space, by learning to translate a textual query into a visual representation allowing text searching in nonannotated (not even automatically) image datasets. Deep Learning is also substituting local features based techniques in smart cameras applications such as parking occupancy detection [6].

While Computer Vision is significantly contributing to make multimedia analysis more effective, the large and increasing amount of multimedia available through social media requires large scale and big data algorithms. Among several approaches to address the problem of efficient search in large archives of image features, one that is very promising is the use of inverted indices. [9] introduced MI-File, an approach that allows using inverted files to perform similarity search with an arbitrary similarity function. In [4,5] a Surrogate Text Representation (STR) derived from the MI-File has been proposed. The conversion of the permutations in a textual form allows using off-the-shelf text search engines for similarity search. Another solution that exploits a text retrieval engine to per-

[4] http://www.deepfeatures.org

form image similarity search, introduced in [7], uses a straightforward quantization of the vector components of the DCNN features. The inverted multi-index uses product quantization both to define the coarse level and for coding residual vectors [14,63]. This approach combined with binary compressed techniques outperforms the state of the art by a large margin [32].

6 Social Mining and Ethics

In a world more and more connected, we are witnessing an incredible growth in the generation and sharing of data originating from the digital breadcrumbs of human activities and sensed as a by-product of the ICT systems that we use everyday. Thanks to the massive availability of this data, human behavior can be observed at large scale. New powerful data-driven tools may be designed and developed to exploit this data for improving the world in many different ways. We can use GPS/GSM data to observe and measure the behavior of a population, to build better cities tailored to the movement of the population, with lower commuting times and lower pollution. We can exploit medical data to build classifiers able to help in diagnosing and curing diseases. We can use industrial data to improve the production processes, and create smarter and more secure factories. We can do a lot of other incredible and useful things with the support of data and analytical tools able to extract useful knowledge from raw data.

These data describing human activities are at the heart of the idea of a *knowledge society*, where the understanding of social phenomena is sustained by the knowledge extracted from the miners of big data across the various social dimensions by using social mining technologies. However, the opportunities of discovering interesting patterns from human data can be outweighed due to the high risks of ethical issues in data processing and analysis and ethical consequences of their suggestions and predictions. Important ethical risks are: (i) *privacy violations*, when uncontrolled intrusion into the personal data of the subjects occurs, and (ii) *discrimination*, when the discovered knowledge is unfairly used in making discriminatory decisions about the (possibly unaware) people who are classified, or profiled.

In the literature some works have shown that data analytics and ethics are not necessary enemies: practical and impactful *data-driven and knowledge-based* services can be designed obtaining data and service quality while enforcing ethical requirements. The key factor is to develop data analytics technologies that *by-design* enforce ethical value requirements to provide safeguards of fairness. This vision is fully compliant with the European General Data Protection Regulation which will be applied on 25 May 2018 and that especially encourages the application of the *privacy by-design* principle.

In the context of privacy protection in big data analytics, Monreale et al. [59] propose the instantiation of the *privacy-by-design* paradigm [23], 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.

Privacy in Data Mining Outsourcing. Giannotti et al. in [45] propose a method for the outsourcing of the association rule mining task while ensuring privacy protection. The results show how an organization can outsource transactional data to an untrusted third party, such as a cloud provider, and obtain a data mining service in a privacy-preserving manner. In this particular scenario, not only the underlying data but also the mined results are not intended for sharing and must remain private because they are considered valuable strategic information. The proposed schema, before sending the transactional data to the third party, applies an encryption based on the addition of fake transactions to the original data in such a way that each item (itemset) becomes indistinguishable with at least $(k-1)$ other items (itemsets). 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. The consequence is that a possible attack has a very limited probability of success in guessing actual items contained either in the transaction data or in the mining results. However, the data owner any time queries the third party can efficiently decrypt correct mining results.

Privacy in Mobility Data Publishing. Monreale et al. [56] present a framework offering an instance of the privacy by-design paradigm concerning personal mobility trajectories, obtained from GPS devices or cell phones. The designed method enforces privacy protection while enabling clustering analysis useful for understanding human mobility behavior in specific urban areas. The released movement data are made anonymous by a process that applies a data-driven spatial generalization of the trajectories. This data-driven approach allows to generalize more areas with high traffic density with respect to urban areas with lower level of traffic. 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 powerful analytic services for info-mobility or location-based services.

Privacy in Distributed Analytical Systems. Monreale et al. in [58] apply the privacy-by-design methodology also in a distributed setting where an untrusted central station is able to collect 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 methodology guarantees for each node of the system privacy protection at individual level by applying a data transformation based on the well-known differential privacy model [33].

In the context of discrimination data analysis, two main lines of research are being pursued (see [68] for a survey).

Discrimination discovery from data consists in the actual discovery of an unjustified difference in treatment of individuals 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

[74]. The process is guided by legally grounded measures of discrimination, possibly including statistical tests of confidence. An alternative view of "discovery as attack" is investigated in [75], in which attack strategies of privacy models are used to unveil discrimination hidden behind redlining practices. Discrimination against individuals has been instead modeled with a k-NN approach, following the legal methodology of situation testing, and applied to a real case study in research project funding [76].
Discrimination prevention consists of removing bias in the machine learning and data mining process. Bias can be present in the training data and in the learning algorithm. Data sanitization for discrimination prevention has been investigated in [73], by first reducing the t-closeness model of privacy to a model for non-discrimination, and then adapting state-of-the- art data sanitization methods for t-closeness. An approach dealing with both privacy and discrimination sanitization is in [49]. Regarding learning algorithms, a modified voting mechanism of rule-based classifiers in order to reduce the weight of possibly discriminatory rules has been proposed [64].

7 Data Infrastructure and Virtual Research Environments as data science enablers

Data Infrastructures [34] open new opportunities to data mining and machine learning activities by facilitating faster and cheaper investigations and enabling a more rapid expansion of their volume. In fact, they are conceived to realise large scale software ecosystems suitable for the big data challenges including analytics [51]. They offer the entire spectrum of resources (data, software, methods, services, computing) needed to carry out a certain investigation "as-a-Service" thus relieving researchers to operate and maintain them. Moreover, they are progressively introducing mechanisms that limit as much as possible the exposure of the researcher to technicalities and challenges related with access to the necessary distributed and heterogeneous set of resources.

Novel data infrastructures support the entire data processing chain, from the collection and preparation of the necessary datasets, to the analytics steps till the publication of the produced outcomes. Along this chain, unifying and open capabilities are provided thus to make it possible, for example, to access uniformly different datasets or to simply plug-in new tools/methods and data whenever needed. The resource space offered can also be made available to researchers through tailored *views*, i.e. web-based working environments known as *Virtual Research Environments* [20], where (a) researchers can focus on a specific investigation by having at their fingerprint what is needed; (b) what researchers produce is equipped with rich enough metadata thus to become a new resource compliant with Open Science [40] practices; (c) researchers are also provided with state-of-the-art facilities promoting collaboration and cooperation.

Driven by this rationale a software system named *gCube* [11] has been designed and developed. This technology is enacting the *D4Science Infrastructure* [19] and exploited to create and operate more than 70 diverse VREs. Overall,

these VREs are serving more than 3100 (returning) scientists in 44 countries across a rich array of diverse communities including the KDD community via the SoBigData RI.

Along the years, gCube has been progressively endowed with (*a*) a rich array of *mediators* for interfacing with existing *systems* and their enabling technologies including distributed computing infrastructures (e.g. EGI [22]) and data providers (e.g. by relying on standards like OAI-PMH, SDMX, OGC W*S) as well as for making it possible for *third-party* service providers to easily exploit gCube facilities (e.g. OAuth, OGC W*S, REST APIs), (*b*) a set of basic services including a shared workspace where the objects used and resulting from VRE activity (beyond simple files) can be stored, organised and accessed as if they were in a "standard" file-manager; a social networking area where the member of each VRE can have discussions, share news and other material of interest, rate each item of a discussion, classify the discussion items by hashtags, refer to people or groups thus to call for actions from them, etc.; a user management area where authorized people are allowed to manage VRE membership, to create groups, assign members to groups, assign roles to member, invite new members, etc.; an open, customizable and extensible set of facilities made available for the needs of the specific community context. These include a project management and issue-tracking system with a wiki, a rich and extensible data analytics platform [11,24], a flexible "products" catalogue where any (research) artefact produced in the VRE and worth being "published" can be easily made available by equipping it with rich metadata including license and provenance-related ones, a rich array of domain data management facilities.

VREs are created by using a wizard-based approach where a VRE designer is simply requested to select (among the available ones) the facilities and resources he/she is willing to have in the VRE, and then upon approval the VRE is automatically provisioned and made available by a web-based portal.

Overall, the data analytics platform resulting from gCube is characterised by the following key principles:

- *Extensibility*: the platform is "open" with respect to (*i*) the analytics techniques and methods it offers and supports and (*ii*) the computing infrastructures and solutions it relies on to enact the processing tasks. It is based on a plug-in architecture to support adding new algorithms and methods as well as new computing platforms;
- *Distributed processing*: the platform is conceived to execute processing tasks by relying on "local engines" / "workers" that can be deployed in multiple instances and execute tasks in parallel and seamlessly. The platform is able to rely on computing resources offered by both well-known e-Infrastructures (e.g. EGI) as well as resources made available by the Research Infrastructures or communities to deploy instances of the "local engines" / "workers". This is key to make it possible to "move" the computation close to the data;
- *Multiple interfaces*: the platform offers its services via both a (web-based) graphical user interface and a (web-based) programmatic interface thus to enlarge the possible application contexts. For instance, having a proper API

facilitates the development of components capable to execute processing tasks from well-known applications (e.g. R, KNIME);
- *Cater for scientific workflows* [54]: the platform is both exploitable by existing WFMS (e.g. a node of a workflow can be the execution of a task / method offered by the platform) and support the execution of a workflow specification (e.g. by relying on one or more instances of WFMSs);
- *Easy to use*: the platform should is easy to use for both (*a*) algorithms / methods providers, i.e. scientists and practitioners called to realise processing methods of interest for the specific community, and (*b*) algorithms / methods users, i.e. scientists and practitioners called to exploit existing methods to analyse certain datasets;
- *Open science friendly*: the platform is transparently injecting open science practices in the processing tasks executed through it. This includes mechanisms for capturing and producing "provenance records" out of any computing task, mechanisms aiming at producing "research objects" so as to make it possible for others to repeat the task and reproduce the experiment.

These key principles make this analytics platform suitable for the challenges KDD community is facing.

8 Conclusions

Twenty-five years ago, most statisticians and computer scientists looked with skepticism at the novel community of KDD scientists, trying to reformulate the analytical process as data driven discovery. Indeed, such visionary endeavor, combined with the advent of big data and spectacular advances in high performance computing, has brought what we call today *data science*: a disruptive paradigm shift impacting all disciplines that pushes towards novel scientific methods where "top down" modelling of phenomena coexists with "bottom up" discoveries from data.

Data abundance combined with powerful data science techniques has the potential to dramatically improve our lives by enabling new services and products, while improving their efficiency and quality.

Many of today's scientific discoveries are already fueled by developments in statistics, data mining, machine learning, network science, databases, and visualization, and we can expect advances in any field related to the comprehension of complex phenomena as in medicine and health (network/personalized medicine), manufacturing (industry 4.0), social dynamics, urban planning, sustainable development.

The importance of data science is widely acknowledged, but there are also great concerns about the irresponsible use of data and models. Automated data driven decisions may be unfair or non-transparent. Confidential data may be shared unintentionally or abused by third parties. Each step in the data science pipeline (from raw data to conclusions) may create inaccuracies, e.g., if the data used to learn a model reflects existing social biases, the algorithm is likely to incorporate these biases. The ethics of data science is a challenging research topic

where computer scientists play a central role, we contributed to change society and we cannot escape the responsibility to understand the impact of the digital transformation helping in catching the opportunities mitigating the risks.

Finally, there is an urgent need to start harnessing these opportunities for scientific advancement and for the social good, compared to the currently prevalent exploitation of big data for commercial purposes (e.g. user profiling and behavioral advertising) or, worse, social control and surveillance. The main obstacle to this accomplishment, besides the scarcity of data scientists, is the lack of a large-scale open ecosystem where big data and social mining research can be carried out.

This is why we propose to establish SoBigData, the Social Mining & Big Data Ecosystem: a research infrastructure (RI) providing an integrated ecosystem for ethic-sensitive scientific discoveries and advanced applications of social data mining on the various dimensions of social life, as recorded by "big data". The research community will use the SoBigData RI facilities as a "secure digital wind-tunnel" for large-scale social data analysis and simulation experiments. SoBigData will serve the wide cross-disciplinary community of data scientists, i.e., researchers studying all aspects of societal complexity from a data- and model-driven perspective, including data and text miners, computer scientists, socio-economic scientists, network scientists, political scientists, humanities researchers, and more.

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: Acm sigmod record. vol. 22, pp. 207–216. ACM (1993)
2. Agrawal, R., Srikant, R.: Algorithms for mining association rules in large databases. In: Proceedings of the 20th VLDB Conference Santiago, Chile. vol. 2, pp. 141–182 (1994)
3. Aliprandi, C., De Luca, A.E., Di Pietro, G., Raffaelli, M., Gazzè, D., La Polla, M.N., Marchetti, A., Tesconi, M.: Caper: Crawling and analysing facebook for intelligence purposes. In: Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on. pp. 665–669. IEEE (2014)
4. Amato, G., Bolettieri, P., Falchi, F., Gennaro, C., Rabitti, F.: Combining local and global visual feature similarity using a text search engine. In: Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on. pp. 49–54. IEEE (2011)
5. Amato, G., Bolettieri, P., Falchi, F., Gennaro, C., Vadicamo, L.: Using apache lucene to search vector of locally aggregated descriptors. In: Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016). pp. 383–392 (2016)
6. Amato, G., Carrara, F., Falchi, F., Gennaro, C., Meghini, C., Vairo, C.: Deep learning for decentralized parking lot occupancy detection. Expert Systems with Applications 72, 327 – 334 (2017)
7. Amato, G., Debole, F., Falchi, F., Gennaro, C., Rabitti, F.: Large scale indexing and searching deep convolutional neural network features. In: International Conference on Big Data Analytics and Knowledge Discovery. pp. 213–224. Springer (2016)

8. Amato, G., Falchi, F., Gennaro, C., Rabitti, F.: Yfcc100m-hnfc6: A large-scale deep features benchmark for similarity search. In: International Conference on Similarity Search and Applications. pp. 196–209. Springer (2016)
9. Amato, G., Gennaro, C., Savino, P.: Mi-file: using inverted files for scalable approximate similarity search. Multimedia tools and applications 71(3), 1333–1362 (2014)
10. Andrienko, G., Andrienko, N., Rinzivillo, S., Nanni, M., Pedreschi, D., Giannotti, F.: Interactive Visual Clustering of Large Collections of Trajectories. VAST: Symposium on Visual Analytics Science and Technology (2009)
11. Assante, M., Candela, L., Castelli, D., Coro, G., Lelii, L., Pagano, P.: Virtual research environments as-a-service by gcube. PeerJ Preprints (2016)
12. Avvenuti, M., Cresci, S., Del Vigna, F., Tesconi, M.: Impromptu crisis mapping to prioritize emergency response. Computer 49(5), 28–37 (2016)
13. Avvenuti, M., Cresci, S., Marchetti, A., Meletti, C., Tesconi, M.: Ears (earthquake alert and report system): a real time decision support system for earthquake crisis management. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1749–1758. ACM (2014)
14. Babenko, A., Lempitsky, V.: The inverted multi-index. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 3069–3076. IEEE (2012)
15. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the 7th Conference on Language Resources and Evaluation (LREC 2010). Valletta, MT (2010)
16. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. science 286(5439), 509–512 (1999)
17. Berlingerio, M., Coscia, M., Giannotti, F., Monreale, A., Pedreschi, D.: Multidimensional networks: foundations of structural analysis. World Wide Web 16(5-6), 567–593 (2013)
18. Bolettieri, P., Esuli, A., Falchi, F., Lucchese, C., Perego, R., Piccioli, T., Rabitti, F.: Cophir: a test collection for content-based image retrieval. arXiv preprint arXiv:0905.4627 (2009)
19. Candela, L., Castelli, D., Manzi, A., Pagano, P.: Realising Virtual Research Environments by Hybrid Data Infrastructures: the D4Science Experience. In: International Symposium on Grids and Clouds (ISGC) 2014 23-28 March 2014, Academia Sinica, Taipei, Taiwan, PoS(ISGC2014)022. Proceedings of Science (2014)
20. Candela, L., Castelli, D., Pagano, P.: Virtual research environments: an overview and a research agenda. Data Science Journal 12, GRDI75–GRDI81 (2013)
21. Carrara, F., Esuli, A., Fagni, T., Falchi, F., Fernández, A.M.: Picture it in your mind: Generating high level visual representations from textual descriptions. arXiv preprint arXiv:1606.07287 (2016)
22. Fernández-del Castillo, E., Scardaci, D., García, Á.L.: The EGI federated cloud e-infrastructure. Procedia Computer Science - 1st International Conference on Cloud Forward: From Distributed to Complete Computing 68 (2015)
23. Cavoukian, A.: Privacy design principles for an integrated justice system - working paper. In: www.ipc.on.ca/index.asp?layid=86&fid1=318 (2000)
24. Coro, G., Candela, L., Pagano, P., Italiano, A., Liccardo, L.: Parallelizing the execution of native data mining algorithms for computational biology. Concurrency and Computation: Practice and Experience 27(17) (2015)

25. Coscia, M., Giannotti, F., Pedreschi, D.: A classification for community discovery methods in complex networks. Statistical Analysis and Data Mining 4(5), 512–546 (2011)
26. Coscia, M., Rinzivillo, S., Giannotti, F., Pedreschi, D.: Optimal spatial resolution for the analysis of human mobility. In: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012). pp. 248–252. IEEE Computer Society (2012)
27. Coscia, M., Rossetti, G., Giannotti, F., Pedreschi, D.: Demon: a local-first discovery method for overlapping communities. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 615–623. ACM (2012)
28. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., Tesconi, M.: Fame for sale: efficient detection of fake twitter followers. Decision Support Systems 80, 56–71 (2015)
29. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., Tesconi, M.: Dna-inspired online behavioral modeling and its application to spambot detection. IEEE Intelligent Systems 31(5), 58–64 (2016)
30. Da San Martino, G., Gao, W., Sebastiani, F.: Ordinal text quantification. In: Proceedings of the 39th ACM Conference on Research and Development in Information Retrieval (SIGIR 2016). pp. 937–940. Pisa, IT (2016)
31. Del Vigna, F., Petrocchi, M., Tommasi, A., Zavattari, C., Tesconi, M.: Semi-supervised knowledge extraction for detection of drugs and their effects. In: International Conference on Social Informatics. pp. 494–509. Springer (2016)
32. Douze, M., Jégou, H., Perronnin, F.: Polysemous Codes, pp. 785–801. Springer International Publishing, Cham (2016)
33. Dwork, C.: Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) Automata, Languages and Programming, Lecture Notes in Computer Science, vol. 4052, pp. 1–12. Springer Berlin Heidelberg (2006), http://dx.doi.org/10.1007/11787006_1
34. Edwards, P.N., Jackson, S.J., Bowker, G.C., Knobel, C.P.: Understanding infrastructure: Dynamics, tensions, and design. Working paper, National Science Foundation (2007), http://hdl.handle.net/2027.42/49353
35. Esuli, A., Sebastiani, F.: Determining term subjectivity and term orientation for opinion mining. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006). pp. 193–200. Trento, IT (2006)
36. Esuli, A., Sebastiani, F.: Sentiment quantification. In: IEEE Intelligent Systems 25(4), 72–75 (2010)
37. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006). pp. 417–422. Genova, IT (2006)
38. Esuli, A., Sebastiani, F.: Determining the semantic orientation of terms through gloss analysis. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM 2005). pp. 617–624. Bremen, DE (2005)
39. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R.: Advances in knowledge discovery and data mining, vol. 21. AAAI press Menlo Park (1996)
40. Fecher, B., Friesike, S.: Open science: One term, five schools of thought. In: Bartling, S., Friesike, S. (eds.) Opening Science, pp. 17–47. Springer International Publishing (2014)

41. Furletti, B., Gabrielli, L., Renso, C., Rinzivillo, S.: Analysis of gsm calls data for understanding user mobility behavior. Santa Clara, California (2013)
42. Gabrielli, L., Furletti, B., Trasarti, R., Giannotti, F., Pedreschi, D.: City users' classification with mobile phone data. In: IEEE Big Data. Santa Clara (CA) - USA (11/2015 2015)
43. Gao, W., Sebastiani, F.: From classification to quantification in tweet sentiment analysis. In: Social Network Analysis and Mining 6(19), 1–22 (2016)
44. Gao, W., Sebastiani, F.: Tweet sentiment: From classification to quantification. In: Proceedings of the 7th International Conference on Advances in Social Network Analysis and Mining (ASONAM 2015). pp. 97–104. Paris, FR (2015)
45. Giannotti, F., Lakshmanan, L.V.S., Monreale, A., Pedreschi, D., Wang, W.H.: Privacy-preserving mining of association rules from outsourced transaction databases. IEEE Systems Journal 7(3), 385–395 (2013)
46. Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., Trasarti, R.: Unveiling the complexity of human mobility by querying and mining massive trajectory data. The VLDB Journal 20(5), 695–719 (Oct 2011)
47. Giannotti, F., Nanni, M., Pinelli, F., Pedreschi, D.: Trajectory pattern mining. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 330–339. KDD '07, ACM, New York, NY, USA (2007)
48. Guidotti, R., Nanni, M., Rinzivillo, S., Pedreschi, D., Giannotti, F.: Never drive alone: Boosting carpooling with network analysis. Information Systems (2016)
49. Hajian, S., Domingo-Ferrer, J., Monreale, A., Pedreschi, D., Giannotti, F.: Discrimination- and privacy-aware patterns. Data Min. Knowl. Discov. 29(6), 1733–1782 (2015)
50. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 675–678. ACM (2014)
51. Khalifa, S., Elshater, Y., Sundaravarathan, K., Bhat, A., Martin, P., Imam, F., Rope, D., Mcroberts, M., Statchuk, C.: The six pillars for building big data analytics ecosystems. ACM Computing Surveys 49(2) (2016)
52. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436–444 (2015)
53. gil Lee, J., Han, J.: Trajectory clustering: A partition-and-group framework. In: In SIGMOD. pp. 593–604 (2007)
54. Liew, C.S., Atkinson, M.P., Galea, M., Ang, T.F., Martin, P., Hemert, J.I.V.: Scientific workflows: Moving across paradigms. ACM Computing Surveys 49(4) (2016)
55. Milli, L., Monreale, A., Rossetti, G., Pedreschi, D., Giannotti, F., Sebastiani, F.: Quantification in social networks. In: Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on. pp. 1–10. IEEE (2015)
56. Monreale, A., Andrienko, G.L., Andrienko, N.V., Giannotti, F., Pedreschi, D., Rinzivillo, S., Wrobel, S.: Movement data anonymity through generalization. TDP 3(2), 91–121 (2010)
57. Monreale, A., Pinelli, F., Trasarti, R., Giannotti, F.: Wherenext: a location predictor on trajectory pattern mining. In: 15th ACM SIGKDD Conference on Knoledge Discovery and Data Mining (KDD'09) (2009)

58. Monreale, A., Wang, W.H., Pratesi, F., Rinzivillo, S., Pedreschi, D., Andrienko, G., Andrienko, N.: Privacy-preserving distributed movement data aggregation. In: AGILE. Springer (2013)
59. Monreale, Anna, Rinzivillo, Salvatore, Pratesi, Francesca, Giannotti, Fosca, Pedreschi, Dino: Privacy-by-design in big data analytics and social mining. EPJ Data Sci. 3(1), 10 (2014), http://dx.doi.org/10.1140/epjds/s13688-014-0010-4
60. Moreo Fernández, A., Esuli, A., Sebastiani, F.: Distributional correspondence indexing for cross-lingual and cross-domain sentiment classification. In: Journal of Artificial Intelligence Research 55, 131–163 (2016)
61. Pappalardo, L., Rossetti, G., Pedreschi, D.: " how well do we know each other?" detecting tie strength in multidimensional social networks. In: Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on. pp. 1040–1045. IEEE (2012)
62. Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., Barabasi, A.L.: Returners and explorers dichotomy in human mobility. Nat Commun 6 (09 2015), http://dx.doi.org/10.1038/ncomms9166
63. Paulevé, L., Jégou, H., Amsaleg, L.: Locality sensitive hashing: A comparison of hash function types and querying mechanisms. Pattern Recognition Letters 31(11), 1348–1358 (2010)
64. Pedreschi, D., Ruggieri, S., Turini, F.: Measuring discrimination in socially-sensitive decision records. In: Proc. of the SIAM Int. Conf. on Data Mining (SDM 2009). pp. 581–592. SIAM (2009)
65. Quinlan, J.R.: C4. 5: programs for machine learning. Elsevier (2014)
66. Rinzivillo, S., Gabrielli, L., Nanni, M., Pappalardo, L., Pedreschi, D., Giannotti, F.: The purpose of motion: Learning activities from individual mobility networks. In: International Conference on Data Science and Advanced Analytics, DSAA 2014, Shanghai, China, October 30 - November 1, 2014 (2014), http://dx.doi.org/10.1109/DSAA.2014.7058090
67. Rinzivillo, S., Mainardi, S., Pezzoni, F., Coscia, M., Pedreschi, D., Giannotti, F.: Discovering the geographical borders of human mobility. KI-Künstliche Intelligenz 26(3), 253–260 (2012)
68. Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review 29(5), 582–638 (2014)
69. Rossetti, G., Berlingerio, M., Giannotti, F.: Scalable link prediction on multidimensional networks. In: Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on. pp. 979–986. IEEE (2011)
70. Rossetti, G., Guidotti, R., Miliou, I., Pedreschi, D., Giannotti, F.: A supervised approach for intra-/inter-community interaction prediction in dynamic social networks. Social Network Analysis and Mining 6(86) (2016)
71. Rossetti, G., Pappalardo, L., Kikas, R., Pedreschi, D., Giannotti, F., Dumas, M.: Homophilic network decomposition: a community-centric analysis of online social services. Social Network Analysis and Mining Journal 6(103) (2016)
72. Rossetti, G., Pappalardo, L., Pedreschi, D., Giannotti, F.: Tiles: an online algorithm for community discovery in dynamic social networks. Machine Learning pp. 1–29 (2016)
73. Ruggieri, S.: Using t-closeness anonymity to control for non-discrimination. Transactions on Data Privacy 7(2), 99–129 (2014)
74. Ruggieri, S., Pedreschi, D., Turini, F.: Data mining for discrimination discovery. ACM Trans. on Knowledge Discovery from Data 4(2), Article 9 (2010)

75. Ruggieri, S., Hajian, S., Kamiran, F., Zhang, X.: Anti-discrimination analysis using privacy attack strategies. In: Proc. of Machine Learning and Knowledge Discovery in Databases (ECML-PKDD 2016) Part II. LNCS, vol. 8725, pp. 694–710. Springer (2014)
76. Ruggieri, S., Turini, F.: A KDD process for discrimination discovery. In: Proc. of Machine Learning and Knowledge Discovery in Databases (ECML-PKDD 2016) Part III. LNCS, vol. 9853, pp. 249–253. Springer (2016)
77. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM 59(2), 64–73 (2016)
78. Trasarti, R., Guidotti, R., Monreale, A., Giannotti, F.: Myway: Location prediction via mobility profiling. Information Systems (2015)
79. Trasarti, R., Pinelli, F., Nanni, M., Giannotti, F.: Mining mobility user profiles for car pooling. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1190–1198. KDD '11, ACM, New York, NY, USA (2011)