Final Remarks on Big Data Analysis and Its Impact on Society and Science

Jerzy Stefanowski and Nathalie Japkowicz

Abstract In this chapter, we summarize the lessons learned from the contributions to this book, add some of the important points regarding the current state of the art in Big Data Analysis that have not been discussed at length in the contributions per se, but are worth being aware of, and conclude with a discussion of the influence that Stan Matwin has had throughout the years on the successive related fields of Machine Learning, Data Mining and Big Data Analysis.

1 Introduction

Big Data is one of the most popular phrases in the current computer science literature. Researchers, specialists working on various applications, philosophers of science and journalists argue that we are living in a new era of the information technology, where Big Data analysis will play a critical role and may change our lives and society.

The rapid development of computer and electronic technology facilitates collecting and processing huge amounts of data. It should be noticed that standard data bases and business transaction systems are not the only, or even the main, sources of such data. Nowadays more and more Big Data is acquired from the Internet, and in particular from social media and other services tracking users' activities. Manufacturing systems, smart meters, sensor and network applications also produce massive volumes of data about business processes, technical conditions of device components, etc. In scientific or engineering tasks, data may have an even more complex structure than the typical business data considered in standard information systems. Big Data manifests itself further in healthcare and medical information systems, which are also a rapidly growing area which includes quite large, complex and heterogeneous data

J. Stefanowski (✉)
Institute of Computing Sciences, Poznań University of Technology, Poznań, Poland
e-mail: Jerzy.Stefanowski@cs.put.poznan.pl

N. Japkowicz
School of Electrical Engineering & Computer Science,
University of Ottawa, Ottawa, Canada
e-mail: nat@site.uottawa.ca

N. Japkowicz and J. Stefanowski (eds.), *Big Data Analysis: New Algorithms for a New Society*, Studies in Big Data 16, DOI 10.1007/978-3-319-26989-4_13

repositories about patients and their treatment. Furthermore, new smart-phones and other mobile devices offer multiple options for data acquisition as well as a variety of data formats, e.g., geographical localization of the phone, records of audio, photographs or other multimedia by internal cameras, the recording of human gestures or movements via accelerators and sensors [37]. This leads toward the development of new software systems, which need to operate efficiently with terabytes of data, often in real time and with tight demands for memory or other resources.

With these new data sources, the machines or measurement devices continuously produce data. Then, the data are processed and analysed by algorithms making human inspection very limited as compared to its role in traditional data analysis. While mining such data may allow us to discover new types of knowledge about people or events, it also opens the gate to new dangers or risks, such as privacy breeches, data protection failures, or ethical issues related to the use of automatic decisions made on human lives or others. These questions were considered that deeply before the advent of Big Data.

The term Big Data does not refer to the massive size or volume only. In the introduction to this book we have briefly surveyed popular definitions of Big Data and stressed the role of other properties—which are often called the "many V's" (besides volume, we also have velocity, variety, veracity, value and variability among others). Therefore, the term "Big Data" encompasses the presence of many heterogeneous data representations, various data sources linked together, complex structures as well as the need to deal with high speed of arrival, processing time requirements, evolving data characteristics, and uncertainty of the data elements. Larger, complex or non-static and, generally speaking, "more difficult" data sets pose new challenges for researchers and call for a variety of new dedicated approaches.

Changes resulting from Big Data are also visible in the new kinds of applications being considered. For example, some researchers attempt to predict an epidemic outbreak based on analyzing web search queries or users' tweets mentioning special-related keywords (see, e.g., a case study of flu prediction [18, 29]). Some other companies identify financial trends by linking many various data sources (in addition to the contents of data bases, social networks, tweets, logs of other Web users' activities and sentiments of their opinions are integrated together and explored [45]); others look for patterns of human mobility by analysing the records of mobile phone calls [16] (note that combining mobile data with additional data may, not only support controlling transport systems, but can also be used by police to predict crime sub-areas for patrols [49]); still others try to optimize the maintenance of city infrastructure or predict dangerous failures of technical devices [40]; some support preparing multi-dimensional astronomical maps by generalizing the results of exploring a huge collection of images covering the sky (see the Sloan Sky Digital Survey [3]); some apply sophisticated, powerful computer technology and natural language algorithms to understand queries and imitate human answers (see experiences with question answering systems and IBM Watson [28]). Finally, e-commerce companies often analyze customers' purchases and track their behavior in order to make more accurate product recommendations, and so on.

We refer the reader to popular books such as [36] to learn about additional successful applications of Big Data Analysis in medicine, natural sciences, engineering and many other fields. These and other case studies described by the authors of this edited book clearly show that Big Data algorithms already have an impact on our daily life and may affect society even more deeply in the near and more distant future.

Machine learning and data mining are among the core methods used in Big Data Analysis. In the introductory chapter of this book we have shown that machine learning researchers had already faced some issues related to the mining of massive and complex data. However, we have also identified several differences between earlier machine learning and present Big Data characteristics of the basic issues (see Sect. 1.2 of chapter "A Machine Learning Perspective on Big Data Analysis" in this book). Besides requirements of computational efficiency, Big Data has opened new research problems, which have never been considered, or considered only in a limited range before. We should also note that the new applications and research challenges can be viewed as multi-disciplinary problems that should be handled by teams of researchers coming from different fields such as databases, data mining, machine learning, statistics, pattern recognition and distributed / parallel computations.

In view of the above considerations and in honour of Stan Matwin who has made a significant contribution to the field over the years,we have decided to prepare this special edited volume. We have invited several well known researchers coming from machine learning and other related disciplines to present their views on how studying Big Data affects the research in their field, to discuss the most interesting new research directions that emerged from their work and to express their opinions and the lessons they have gathered from their experience. Furthermore, we have encouraged them to discuss the impact of Big Data on society as well as the possible dangers or risks that such research could cause.

In the next section we summarize the main problems raised by our authors and describe some of the lessons learned from the case studies discussed in their chapters. Then, in Sect. 3, we briefly survey some other Big Data opportunities and challenges that were not considered in great detail by our authors, but in our opinion, are important to consider given their influence on Big Data analysis research and on society in general. The final section of this chapter concludes the book by presenting a short summary of Stan Matwin's influence on the field of Big Data Analytics.

2 Lessons on Big Data Opportunities and Challenges

In this section, we will describe several issues or problems raised by our authors and conclude with some promising research directions they pointed out. Please note that this section is organized by themes which are conveyed by our subsection titles. Many of our authors' contributions raise several of these themes simultaneously and will, consequently, be discussed in several of our subsections.

2.1 Relation Between Traditional Data Analysis and Big Data Research Paradigms

Many popular texts on Big Data talk about changes in the way scientific research will increasingly be carried out. They suggest that sophisticated statistical models will be replaced by more data intensive methods applied to huge amounts of data. Similarly, controversial opinions are expressed concerning claims that discovering correlations in big data sets reduces the needs for discovering causality and attempts at understanding the data. The question of whether Big Data offers a new less theoretical methodology has also been raised in a few chapters of our book.

R. Sparks, A. Ickowicz and H. Lenz discuss such questions in chapter "An Insight on Big Data Analytics", within the historical context of modern science and in particular, statistical analysis. They overview the historical traditions of statistics, which was often based on constructing models to better understand the data. However, they also explain that statisticians use empirical models to approximate "real data models", and integrate this with mathematical theory to understand processes and construct knowledge. Although traditional statistical methods are based on strong theoretical assumptions and intuitive models—due to the small size of data samples and the strict requirements imposed on them—statisticians recognize that some of these theoretical frameworks are too restrictive. They say that such a realization is a useful step in the right direction for finding new ideas to solve problems without making unrealistic assumptions. On the one hand, R. Sparks, A. Ickowicz and H. Lenz cite works of philosophers of science, such as I. Kant or C. Popper that suggest it is impossible "to start with pure observations alone, without anything in the nature of a theory". On the other hand, they also conclude that "good models shape the data in trying to best fit it, and that the data also shapes the model in that it helps to use models with the appropriate assumptions".

In their chapter R. Sparks, A. Ickowicz and H. Lenz also nicely describe the tension between the traditional statistics and data mining communities. Data mining is usually seen as a methodology that favours data-exploration at the expense of theory. Moreover, in the authors' opinion, non-statistically trained data-miners quite often drop theoretical considerations and test a lot of methods. This, they believe, is an unsound approach. The authors thus disagree that Big Data is going to drive knowledge in the complete absence of a theoretical framework or models.

They also postulate that data-miners and statisticians should collaborate more closely in mining big data sets and in generating knowledge within a sound theoretical framework. They believe that statisticians should stop making unrealistic assumptions that remain unchecked, and that data-miners should work with statisticians in helping discover sound knowledge that will help manage the future. They also ask the question of whether data mining methods may help construct an appropriate empirical model. They argue that statisticians need to be more pragmatic and nicely refer to Breiman's paper on two statistical cultures [4], which discusses a kind of shift from model driven approaches to algorithm modeling-based approaches. Finally, they list examples of successful non-parametric methods, such as ensembles

or Bayesian approaches, that have been developed by statisticians in the last few decades and represent this new methodological paradigm, and are, simultaneously, of particular interest to data miners.

An interesting integration of statistical and machine learning approaches is presented in chapter "Discovering Networks of Interdependent Features in High- Dimensional Problems" by M. Draminski, M. Dabrowski, K. Diamanti, J. Koronacki and J.Komorowski, where methods for feature selection taking into consideration feature interdependencies in genomic data are developed.

Finally, we direct the reader to Sect. 2 in chapter "A Machine Learning Perspective on Big Data Analysis" of our introductory chapter where we survey the literature on the above issues and present different arguments for and against the claim that a Big Data revolution is underway. In particular, regarding the role of statistics in Big Data Analysis, we summarize, in Sect. 2 of chapter "A Machine Learning Perspective on Big Data Analysis", the arguments of several researchers concerning sample and selection biases that cannot be eliminated, illusions of working with the complete population, unknowns in the data, needs to understand causality, and the fact that correlations are not always sufficient to take actions in the real world.

2.2 The Need to Develop New Methodological Frameworks for Complex Problems

The call for the development of new methodological frameworks in the context of distributed data sets and complex interaction systems is expressed by A. Skowron, A. Jankowski and S. Dutta in chapter "Toward Problem Solving Support Based on Big Data and Domain Knowledge: Interactive Granular Computing and Adaptive Judgement". These authors notice that big data sets are often distributed and their parts are linked together. Moreover, computations are performed on quite complex objects and often affected by uncertainty. They argue that such computations are distributed over networks of agents involved in complex interactions. Agents perform computations on complex objects of very different natures, e.g., (behavioral) patterns, classifiers, clusters, structural objects, sets of rules, aggregation operations, reasoning schemes, etc. Moreover, implementing the process of mining such complex data sets in distributed networks is related to modeling the complex analytical systems with some basic high level primitives for composing and building complex analytical pipelines over Big Data. Such primitives are very often expressed in natural language, and they should be approximated using other low-level primitives, accessible from raw data or from domain expert knowledge.

To model such complex systems at the higher level, A. Skowron, A. Jankowski and S. Dutta propose to exploit the paradigm of Granular Computing. Granulation of information should be considered when precision of information is too costly and not very meaningful in modeling and controlling complex systems. Moreover, data are incomplete, uncertain, and vague. The granular computing paradigm is based

on soft computing, such as fuzzy or rough sets theory. Although basics of Granular Computing have already been proposed, also by A. Skowron in his earlier papers [44], here authors extend it by introducing new complex granules. They also show how to build such granules using data and approximating expert's ontologies, in particular for hierarchical and multi-domain approaches. Moreover, they present a new unified methodology for modeling and controlling computations with these complex granules in case of an interaction between agents. They argue that it could support users in solving problems of Big Data, as granules may represent computational building blocks for approximating (or inducing models of) the high-level primitives used by researchers to compose complex analytical pipelines over Big Data.

2.3 Dynamic and Evolving Data

Data streams are one of the most challenging forms of Big Data, in particular if data evolve over time. In supervised machine learning, such unexpected changes in the underlying data distribution over time are referred to as concept drift. Such changes deteriorate the predictive accuracy of classifiers learned from past examples and they require new learning algorithms that could detect and adapt to concept drifts.

I. Zliobaite, M. Pechenizkiy and J. Gama provide an application-oriented overview of research in this field. The original contribution of their chapter "An Overview of Concept Drift Applications" includes a detailed survey of concept drift handling methods and focuses the reader's attention to the new research tasks driven by the typical categories of applications found in the context of data streams.

First, the authors overview and categorize application tasks for which the problem of concept drift is particularly relevant. Then, they introduce a special reference framework for describing application-oriented tasks in a systematic way. Their original proposal for the framework includes three main components:

1. The main properties of the application tasks with concept drift (data and learning tasks, characterization of changes and operational setting for availability of the ground truth as class labels).
2. A categorization of application areas and tasks based on those properties (they distinguish mainly between monitoring and control, information management, and analytics and diagnostics applications).
3. Links between tasks and applications.

The authors noticed that their categories of applications differ in terms of the data types they use. Monitoring and control applications typically use streaming sensory data as inputs, and concept drifts typically happen fast and suddenly. Information management applications work with documents and concept drifts happen more slowly than in the previous case. Diagnostic applications typically use time-stamped observations and concept drifts are even slower—typically incremental, or evolving. Then, they survey application-oriented published works on adaptive learning and focus on a few application examples that represent different types of tasks. Using

these examples, they illustrate how the prediction task is formulated, and how concept drifts are handled.

The other interesting lesson from this chapter is related to the implications of evolving data on current research in data mining. Although the problem of concept drift has been recognized as an interesting one, the current state of research is still in an early stage and the many proposals that have been formulated were examined in artificial and theoretical settings. Indeed, these approaches have been tested primarily on simulated data or real data with simulated drifts. Assumptions behind expected type of changes and reasons for the changes are not always precisely explained and studied. I. Zliobaite, M. Pechenizkiy and J. Gama postulate that more studies should highlight the peculiarities of particular applications and give intuition and/or empirical evidence as to why traditional general-purpose concept drift handling techniques are not expected to perform well. They also suggest more research on specialized techniques suitable for a particular application type in the real world context.

Their other lessons from several real-life projects are the following: seasonal effects with vague periodicity for a certain subgroup of object occur in some problems, external contextual information (which could be available) or extraction of hidden contexts from the predictive features may help handle recurrent concept drift better, mining temporal relationships can be used to identify related drifts, domain experts should play an important role in acceptance of Big Data solutions. These experts will slowly move from non interpretable black-box models towards control systems that support an understanding of how these changes are detected and what adaptation would happen.

They expect changes in research on concept drift and hope that these changes will be helpful in improving utility, usability and trust in the adaptive learning systems being developed for many of the Big Data applications.

R. Sarmento, M. Oliveira, M. Cordeiro, S. Tabassum and J. Gama also consider the analysis of real-time streaming data in chapter "Social Network Analysis in Streaming Call Graphs". They discuss the challenges encountered in the analysis and visualization of the network data collected by mobile network operators. As the conventional data analysis performed by telecom operators is slow and implies heavy costs in data warehouses, the authors have modeled these time changing graphs as a data stream. This modeling combined with a special sampling has helped the visualization of mobile graphs. To sum up, this chapter nicely illustrates the authors' research in network sampling, visualization of streaming social networks, stream analysis and online exploratory data analysis.

Postulates for developing new types of adaptive learning algorithms that should lead to incremental models from asynchronous streams coming from financial application are also discussed by E. Paquet, H. Viktor and H. Guo in chapter "Data Mining in Finance: Current Advances and Future Challenges". Finally, M. Shah, in chapter "Big Data and the Internet of Things" also argues that the speed and scale at which the smart devices of the Internet of Things produce data require new streaming algorithms.

2.4 Information Network Analysis

Nowadays many of Big Data applications come from the Web, social media, or more generally, from networks. As such, these applications are connected with the analysis of information networks. In chapter "Analysis of Text-Enriched Heterogeneous Information Networks", J. Kral, A. Valmarska, M. Grcar, M. Robnik-Sikonja and N. Lavrac present a brief history of this research starting from sociologists like Zahary and progress toward the current state-of-the-art in network analysis. Issues of graph mining and social network analysis were also surveyed in the introductory chapter of this book.

J. Kral, A. Valmarska, M. Grcar, M. Robnik-Sikonja and N. Lavrac focus the readers attention on heterogeneous information networks [46]. These types of networks describe heterogeneous types of entities and different types of relations. Moreover, in enriched heterogeneous information networks, nodes of certain types contain additional information. The authors introduce us to this newer research area. They also claim that the methods that take the heterogeneous nature of the networks into account are capable of solving tasks that cannot be defined on homogeneous information networks (like clustering two disjoint sets of entities). They show how to merge the network analysis with the analysis of other data formats, either in the form of text documents or results obtained from various past experiments. The novel contribution of the authors chapter is to present a method for mining text-enriched heterogeneous information networks, which combine the information stored in a heterogeneous network with textual data. Unlike the related approaches, the new method combines two separate sources (network structure and text) and joins them into a single representation.

Their chapter also includes two case studies illustrating this method. The results obtained on the VideoLectures.NET data show that using this method increases classification accuracy as compared to using only texts or only structural information about the instances. Moreover, the results obtained by their other study on psychology paper bibliographies show that the relational information hidden in the network structure is particularly useful.

Chapter "Social Network Analysis in Streaming Call Graphs" by R. Sarmento, M. Oliveira, M. Cordeiro, S. Tabassum and J. Gama describes a real life case study of telecommunication data transformed into graphs, where nodes represent subscribers and edges represent the phone calls. The authors discuss which aspects of the analysis of the social networks underlying call graphs can deliver valuable business insights to mobile telecom operators (e.g., topological aspects of the networks in terms of degree distribution, average path length, clustering, connected components and finding the key nodes in the network based on the position they occupy in the network structure, community detection, etc.). They also show other challenges pertaining to the network data collected by mobile network operators: data are more complex, they are continuously generated by the communication activity among subscribers, and in addition to the large volume, this data arrives at high rates. Therefore, the authors point out requirements for developing new methods that should be able to

cope with data speed and volume and operate under the one-pass paradigm. This led them to model these call graphs as data streams, generate specialized sampling for them, as well as new techniques for visualization which were discussed in the previous subsection.

Combinations of various data types are also discussed in chapter "Industrial-Scale AdHoc Risk Analytics Using MapReduce" by E. Paquet, H. Viktor and H. Guo where the advantages of combing financial data analysis with non-traditional data (social, tweets, etc.) and its impact for trend predictions are considered.

2.5 Mining Sensing Data and Exploiting the Internet of Things

Advances in sensing and information technologies are making it possible to embed increasing computing power in small devices and open up new opportunities for both collecting and processing large–scale data. Combined with advances in communication, this results in a system of highly interconnected devices referred to as the Internet of Things. It is claimed that the Internet of Things will be a growing source contributing to Big Data Analysis in the nearest future [37].

In order to mine sensing data, the existing data mining techniques have to be adapted to dealing with constraints in resources and to performing an analysis in real–time. The underlying focus of ubiquitous systems is to perform computationally intensive analysis techniques on mobile device environments that are constrained by limited computational resources and varying network characteristics. Furthermore, it becomes necessary to perform synthesis and knowledge integration from multiple data streams in a resource constrained environment.

These problems are discussed in chapter "Big Data and the Internet of Things", where M. Shah presents several important aspects of the intersection of Big Data analytics and the Internet of Things. The brief review of the connectivity, communication and data acquisition issues is not the main focus of the chapter. Instead, the author focuses on the novel opportunities and challenges that the new world of interconnected devices offer, along with some advancements that are being made on various fronts to realize them.

In this chapter, M. Shah discusses how Big Data technologies and the Internet of Things are playing a transformative role in society. In his view, the ubiquitous nature of such technologies will profoundly change the world as we know it, just as the industrial revolution and the Internet did in the past. He expects that they will change the context in which predictive analytics is performed in many application problems (examples of real-time diagnostics of air-engines and electrical turbines are considered to illustrate this statement). The author predicts that some devices will take corrective actions, thus making themselves self-aware and self-maintaining.

Other lessons from M. Shah's chapter include recommendations for business organizations or companies developing applications at the intersection of Big Data and

the Internet of things. Besides several computational and technological needs, he postulates that it will also become necessary to facilitate interfacing between engineering or domain experts and data scientists for efficient and productive knowledge transfer, agreed-upon validation, as well as adoption and integration mechanisms for analytics.

Finally, he argues that researchers and company have to pay more attention to societal aspects of the new technologies. He lists the following areas that need to receive more attention and efforts than they currently do: privacy, security, and interpretability of models and data quality issues. Other challenges pertain to issues resulting from the difficulties associated with the adoption of data mining analytics in various domains, e.g. the limitations of model validation and testing, the integration of the Internet of things devices with the human physical understanding of the world, the risks of systemic errors and failures.

2.6 The Need to Deal with Heterogeneous Representations, Vagueness and Unknown Data

Variety as it refers to heterogeneous data is one of the essential properties of Big Data. These different data forms are usually greatly interconnected, interrelated and may also be inconsistently represented which creates challenges for their integration and cleaning. Heterogeneity also forces analysts to deal with structured, semi-structured and unstructured data simultaneously, which is another difficult task to approach when using standard knowledge discovery tools.

These issues are discussed in a few chapters of this book. J. Kral, A. Valmarska, M. Grcar, M. Robnik-Sikonja and N. Lavrac in chapter "Analysis of Text-Enriched Heterogeneous Information Networks" present a new method, which combines structural information separately calculated from homogeneous networks with the text vector representation (obtained from textual information contained in network nodes). Their case study illustrates that combining these two different heterogeneous representations is feasible and is more powerful than standard methods that handle them independently.

Dealing with heterogeneous data is a major challenge in the integration phase of knowledge discovery. M. Shah in chapter "Big Data and the Internet of Things" describes the current efforts to standardize data protocols for data exchanges between various measurement devices and computer systems. However, he warns readers that the high resolution and temporal nature of such data makes it difficult to align multiple sources as well as devise strategies to learn from them in conjunction with static data sources. The protocols for obtaining the quantities from different measurement systems are still not uniform or standardized even within a given domain. The data integration becomes more difficult since it requires the transformation of such derived quantities and the solving of many conflict situations.

Moreover, in chapter "An Insight on Big Data Analytics", R. Sparks, A. Ickowicz and H. Lenz discuss the usefulness of statistical tools for integrating and reducing large data sets. The complexity of basic data elements, their vague description and several problems of using imprecise natural language are also mentioned in A. Skowron, A. Jankowski and S. Dutta in chapter "Toward Problem Solving Support Based on Big Data and Domain Knowledge: Interactive Granular Computing and Adaptive Judgement", where they postulate the development of new data mining methods for dealing with such data.

Similarly, E. Paquet, H. Viktor and H. Guo consider unknowns (data, parameters, etc.) associated with financial data. The authors show how analyzing and understanding which attributes and parameters are not known is crucial in order to create accurate and meaningful predictions.

2.7 Process of Knowledge Discovery from Data

Although Big Data projects may concern various data sets and involve quite different techniques, some of our authors suggest that more investigations into the systematic process approach to discovering knowledge and deploying final models are needed. Recall that in the practice of Knowledge Discovery from Data, such a way of thinking has resulted in useful standards, such as the CRISP-DM model [6]. This is also a leitmotif in chapter "Implementing Big Data Analytics Projects in Business" of F. Fogelman-Soulie and W. Lu, where the opportunities created by Big Data analytics for companies and the challenges associated with the practical implementation of such projects are discussed. In their view, the process of implementing Big Data Analysis projects in companies includes a number of stages that were inferred from earlier data mining projects, however, they believe that more efforts need to be put into integrating, cleaning and pre-processing the data.

They also claim that appropriate feature engineering is very meaningful for the business domain since such data sets are often high dimensional. Reports from various business or industrial projects show that working with at least 1,000 features is common, but some projects may generate even more features. However, the feature engineering stage is a very difficult step to perform given that it requires lots of data, large computation time and more complex models.

In Big Data Analysis problems, some additional features can be obtained from outside data sources, such as open data sources or private data obtained from partners or data providers. These new data may bring additional value. However, as they are of different formats and semantics—a problem reflected in the Variety of data issue—they need careful realizations of many transformation steps in pre-processing. Compared to earlier machine learning applications, these steps require new models and software tools. F. Fogelman-Soulié and W. Lu review some open-source tools in the section entitled "Architectures for Big Data" in their chapter. These authors nicely illustrate how feature engineering, especially with different semantics, can increase

the performance of the final model by describing a real project of credit-card fraud detection on the Internet.

M. Draminski, M. Dabrowski, Kl. Diamanti, J. Koronacki and J. Komorowski also consider in chapter "Discovering Networks of Interdependent Features in High-Dimensional Problems" new methods for the identification of the most important and independent features in bio-informatics data. They argue that higher numbers of relevant features may be more challenging to obtain than increasing the number of observations.

An additional issue raised by F. Fogelman-Soulié and W. Lu in other parts of their chapter is that choosing appropriate learning algorithms from the many existing ones is not a trivial task. Like other researchers before them, they suggest that a practical lesson drawn from recent Big Data projects is that simple models with lots of data could perform better than complex models on less data. They propose an incremental strategy where the analyst should choose a relatively simple algorithm and work with increasing data volumes with feature engineering. Simpler algorithms are also easier to explain than more complex ones, so sometimes, domain experts or users will prefer simpler models to more accurate algorithms such as ensembles, due to their better interpretability. Finally, they warn readers of the importance and difficulty of choosing appropriate procedures for evaluating learning algorithms, in particular if bigger data are divided into smaller samples or when data sets are progressively increased (either by adding observation, or features).

Quite similar practical observations on the interpretability and evaluation of proposed models can be found in M. Shah's chapter—see the previous Sect. 2.5 in chapter "Big Data and the Internet of Things".

Finally, I. Zliobaite, M. Pechenizkiy and J. Gama present yet another process approach in chapter "An Overview of Concept Drift Applications". They start by discussing the classical model of the data mining process (the CRISP-DM standard), where the life cycle of a data mining project spans over six phases: business understanding, data understanding, data preparation, modeling, evaluation and deployment. As this model assumes that most of the data mining steps are performed offline, it is not appropriate for data streams. Therefore, they generalize it to the streaming settings, where concept drifts and changes of models are expected. The main differences between their proposed model and the standard process is that the data preparation, mining, and evaluation steps are completely automated, there is no manual data exploration, and there is an automated monitoring of performance, including change detection and alert services.

2.8 Architectural Support for the Efficient Mining of Big Data

Big Data requires new technologies to efficiently process huge amounts of data within a tolerable time. Standard storage disk systems may be too slow and limited for new tasks. Therefore, new storage infrastructures, suitable for parallel processing nodes

have recently been developed [43]. Other technologies commonly applied to Big Data include massively parallel and distributed processing. Real, or near-real, time information processing and delivery of results is one of the requirements for Big Data Analytics in many applications. Massive, evolving and complex data characteristics lead to the development of new scalable algorithms allowing for data processing and analyzing. New architectures (software or hardware) for the efficient management of complex and dynamic data streams and their analysis (sometimes in an approximate way) are required as well.

Many of the chapters in this book consider these issues. They note that standard relational database management may be insufficient for the storage and management of big data sets. For instance, F. Fogelman-Soulié and W. Lu refer to efforts by many companies to integrate various data repositories into data warehouses. They discuss the difficulties and cost of constructing ETL models (i.e. Extraction, Transformation, Load of data into data warehouses) and their implementations in financial companies. However, considerations of heterogeneous representations, dynamic, constantly emerging data sources and other characteristics of Big Data have led them to conclusions that fixed static and structured data warehouse models are not adequate. To cope with these limitations they propose to use a new architecture, called "Data Lake" which is a special repository of all the data collected by an organization, where the data is stored in its original raw form. Because no a-priori structure or data model is imposed at collection time, all further usage should be possible without having to modify a pre-existing model.

Furthermore, M. Shah, in Sect. 2 of chapter "Big Data and the Internet of Things" surveys the problems of data integration and management in the context of mining data from mobile and sensing devices. He also explains why classical relational databases are no longer sufficient for dealing with such diversified data sources. NoSQL databases are the answer to these limitations. He advocates the use of columnar data stores such as BigTable, Cassandra, Hypertable, HBase (inspired by the BigTable); key-value and document databases such as MongoDB, Couchbase server, Dynamo and Cassandra; stream data stores such as Eventstore; graph based data- stores such as Neo4j and so on. A slightly more comprehensive description of these systems is also available in our introductory chapter.

The next issue concerns processing platforms. F. Fogelman-Soulié and W. Lu present a nice historical discussion of the tradeoff between traditional big servers (with a scaling-up mechanism) and clusters of less costly simpler machines.

Other authors of this book refer to the Hadoop distributed file system and MapReduce as solutions for running large-scale distributed Big Data processing applications. M. Szczerba, M. Wiewiórka, M. Okoniewski and H. Rybiński present an overview of cloud-based Big Data analytic tools that are currently used and developed for genomic data analysis and that are based on tools coming from the Hadoop system. M. Shah briefly discusses Hadoop relevance to dealing with data coming from the Internet of Things.

An interesting example of programming in the MapReduce framework is described in chapter "Industrial-Scale Ad Hoc Risk Analytics Using MapReduce" by A. Rau-

Chaplin, Z. Yao, and N. Zeh in the context of performing large- scale Monte Carlo simulations to approximate the portfolio-level in their risk analysis system.

Several other authors also notice the limitations of distributed systems such as Hadoop, with respect to time delays in performing analytics. Many machine learning algorithms require multiple passes on the data that are too costly in terms of communication with the underlying system. They direct our attention to newer frameworks such as Spark that were developed to address these issues and are more suitable for intensive machine learning and data mining scenarios (see, e.g., Sect. 3 of the chapter "Scalable Cloud-Based Data Analysis Software Systems for Big Data from Next Generation Sequencing" by M. Szczerba, M. Wiewiórka, M. Okoniewski and H. Rybiński). Then, F. Fogelman-Soulié and W. Lu describe the use of Spark tools inside the idea of a Big Data platform (see Sect. 5 of chapter "Implementing Big Data Analytics Projects in Business").

2.9 Domain-Specific Cases of Big Data Analysis

The chapters of this book also include the description of several Big Data Analysis applications to various problems. The three dominant application areas considered by the authors are life science (mainly biomedicine and genomics), business (mainly finance) and technology.

Life Science

M Szczerba, M. Wiewiórka, M. Okoniewski and H. Rybiński discuss in chapter "Scalable Cloud-Based Data Analysis Software Systems for Big Data from Next Generation Sequencing" problems of mining sequenced data coming from various molecular biology laboratory technologies (e.g., applications pertaining to DNA genotyping, RNA expression profiling, genome methylation searches, and many others). Due to the decreasing costs of the sequencing machines, the amount of collected biological data has significantly increased. The next generation of sequencing technology should consequently contribute much more to Big Data and will influence new diagnostics in medicine. The results of analyzing genomic data can be used in many stages of diagnosing and treatment procedures, especially for personalized medicine, as well as for constructing new functional knowledge bases. However, it causes challenges for efficient storage and data analysis. Discussing these challenges and dedicated software and architectural solutions are the main contributions of their paper. First, the authors present a very interesting overview of Big Data analytic cloud tools that are currently used, tested or are adapted for genomic data analysis. They describe examples of tools developed on the basis of Hadoop and Spark platforms. Moreover, their chapter gives a detailed case study of a special tool, called SparkSeq. It is the dedicated genomic big data processing system, which has already been applied in a number of biological sequencing analysis projects. Perspectives for similar system applications in biology and medicine are also discussed. The final

sections of this chapter includes the authors view on the next generation sequencing big data architectures and open problems of developing new scalable software tools for bioinformatics.

Genomic applications are also considered in chapter "Discovering Networks of Interdependent Features in High-Dimensional Problems" by M. Draminski, M. Dabrowski, K. Diamanti, J. Koronacki and J. Komorowski. Their new methodology for selecting features and discovering their interactions is validated on a large, fairly complex real data set concerning gene expression levels in some human cells. The authors showed that their Monte-Carlo Feature Selection MCFS-ID algorithm returned a limited number of highly informative features, which could also support learning accurate classifiers. They also showed the usefulness of their other method for constructing Inter Dependent Graphs (for detecting strong interactions between features, and using a special approach to analysing rules discovered from data) on the same kind of the gene expression data set. These graphs and underlying rules provide experts with a refined view of biological results and support their interpretations. To sum up, this chapter shows that new methods for feature engineering are necessary in Life Science (where data sets are often highly dimensional) and the combination of such methods with the construction of graphs of interactions between features may help in understanding complex relations in bio-medical data.

Business and Financial Analysis

A few other authors considered the context of financial or more general economic problems.

For instance, A. Rau-Chaplin, Z. Yao, and N. Zeh discuss problems of risk analysis for reinsurance companies in chapter "Industrial-Scale Ad Hoc Risk Analytics Using MapReduce". They showed that typical systems for aggregate risk analysis are efficient at generating a small set of key portfolio metrics required by rating agencies and other regulatory organizations. However, these systems are not able to deal with ad hoc queries that provide a better view of the many dimensions of risks that can impact a reinsurance portfolio. To ensure better financial planning, the insurance companies need to carry out large-scale Monte Carlo simulations to estimate the probabilities of the losses incurred due to catastrophic or critical events. These more advanced risk-analysis queries and simulations require stronger computing power and are both data-intensive and time demanding. The main contributions of their chapter include: discussing new distributed and parallel solutions for such risk estimation with references to Big Data techniques, and presenting the authors' system which uses the MapReduce framework and carefully engineers data structure implementations.

Chapter Data Mining in Finance: Current Advances and Future Challenges by E. Paquet, H. Viktor, and H. Guo also addresses the issue of making predictions and building trading models for financial institutions. These authors provide a short overview of the current development of Big Data in this sector. Then, they focus on particular characteristics that occur in Big Data sets in the financial sector: unknown values and parameters, and randomness in the financial models. In their opinion,

traditional data mining techniques are too limited to deal with such data characteristics. They describe stochastic predictive models for financial data, Although the major part of chapter "Big Data and the Internet of Things" by M. Shah concerns Big Data and the Internet of Things, the author also discusses many application domains impacted by Big Data analytics. He expects changes in the manufacturing sector, asset and fleet management, operations management, resource exploration, energy sector, healthcare, retail and logistics. Section 3 of chapter "Big Data and the Internet of Things" includes an illustrative case study, and a discussion of the opportunities that may arise from mining Big Data by showing its impact on organizations focusing on these domains. The next sections of this chapter are of great interest as well as they include a discussion of the necessary changes an organization is willing or capable to make in order to implement Big Data projects (see Sect. 4 in chapter "Big Data and the Internet of Things"), and the author's opinion on more general societal impact and areas of concerns (Sect. 5 of chapter "Big Data and the Internet of Things") which should be more appropriate for the high Volume and Variety of Big Data encountered in their area of application. The other part of their interesting discussion concerns the evolving aspect of financial data. These include highly fluctuating data, data arriving at a fast rate, late-arriving data, etc. (see Sect. 6 of chapter "Data Mining in Finance: Current Advances and Future Challenges").

Finally, F. Fogelman-Soulié and W. Lu illustrate their considerations with a real life project of credit-card fraud detection on the Internet, funded by the ANR (the French National Research Agency). This is an important area of applications for new data mining methods. It becomes more critical due to the increases in Internet transactions and in the activity of crime groups. The authors discuss the volume of collected transaction data, the specific limits of the recorded data items and their dynamic characteristics. The important part of their case study is to construct appropriate feature representation and to describe their experiences with building and evaluating good prediction models.

Technological Applications

Although the major part of chapter "Big Data and the Internet of Things" by M. Shah concerns Big Data and the Internet of Things, the author also discusses many application domains impacted by Big Data analytics. He expects changes in the manufacturing sector, asset and fleet management, operations management, resource exploration, energy sector, healthcare, retail and logistics. Section 3 of chapter "Big Data and the Internet of Things" includes an illustrative case study, and a discussion of the opportunities that may arise from mining Big Data by showing its impact on organizations focusing on these domains. The next sections of this chapter are of great interest as well as they include a discussion of the necessary changes an organization is willing or capable to make in order to implement Big Data projects (see Sect. 4); and the authors opinion on more general societal impact and areas of concerns (Sect. 5 of chapter "Big Data and the Internet of Things").

Finally, in chapter "Social Network Analysis in Streaming Call Graphs" R. Sarmento, M. Oliveira, M. Cordeiro, and J. Gama describe some of the problems that are encountered in the particular sector of telecommunications services. Their paper

concerns the analysis of the very large and dynamic telecommunication networks graphs, looking for patterns of interactions between users. The authors also propose innovative visualization techniques and describe their implementation. Results of the analysis of such graphs provide useful insights into the social behaviors of users. These behavioral patterns provide significant gains to telecom service providers, e.g., maximizing profits by customer segmentation, profiling, churn and fraud detection etc. Apart from this, they also provide benefits to society in terms of users or subscribers.

3 Other Research Challenges of Big Data Analytics

In this section we very briefly discuss a few other issues, which have an impact on society and research.

3.1 Privacy and Ownership of Data

Privacy issues have become very important with the advent of Big Data and may have a great societal impact. Stan Matwin, as a matter of fact, is one of the first data mining researchers who have recognized this very dangerous side-effect of learning methods, warned researchers about it and looked for solutions to counter it. He came to that problem from moral and ethical concerns. In his words [33]:

> My interest in data privacy is a little different. I am concerned about the fact that modern computers may become a tool that can be used to breach and violate people's privacy easier and on a much larger scale than it was possible, say, 30 years ago. I believe that since the computer research community invented the tools that make it possible—databases, the internet, image and voice recognition, barcodes, etc.—it is then our moral obligation to at least think about tools that would make privacy easier and that would avoid many privacy-averse incidents

He has been working on developing methods that make it nearly impossible to identify a given individual in a data set [35, 53, 54].

We noticed our authors awareness of these problems as well. For instance, the reader can have a look at Sect. 8 of chapter "An Insight on Big Data Analytics" where the authors asked several important questions concerning the ownership of data sets, confidential agreements, new views on intellectual property of the data, unsolved limits of sharing data sets and integrating them from different sources. Moreover, these authors discuss various consequences of applying data mining results. M. Shah warns, in chapter "Big Data and the Internet of Things", that the current methods for privacy preserving data mining are still at a preliminary phase and that efforts to deal with that issue, to-date, have focused mainly on the data and basic analytics stage. He argues that the Internet of things applications have more specific requirements that should be properly addressed in future research.

Looking more widely in the literature, one can find more opinions saying that we still do not know how to share private data while ensuring that the data remains useful. The current techniques for maintaining privacy are too weak to allow the mining of Big Data with high quality results [9, 39]. It is believed that certain paradigms such as differential privacy reduce the information content too much to be useful in practical situations [50]. At the other extreme, as previously mentioned, data may be adequate for mining algorithms but, in such cases, privacy is not always properly considered.

Another related issue concerns the right of people to their own electronic records, and the understanding that their data is often used for analytical aims other than those they envisioned. The majority of users of on-line systems do not go beyond their basic level of data control, and they do not know what it means to share data or that their data (even web search phrases) will be linked to other data sources and mined to provide new results. Yet another ethical problem is using the results of mining personal data to predict the actions of other people.

All these and other issues open up many additional challenging problems. Some of them are more algorithm–oriented, while others are open law questions. Teen and Polonetsky call for new models balancing benefit for researchers and individual privacy rights [47]. As suggested in [38] the "foundations of data mining need to be reformulated in such a way that privacy protection and discrimination prevention are embedded in the foundations themselves, dealing with every moment in the data-knowledge life cycle, from data capture to data mining and analytics, up to the deployment of the extracted models".

3.2 Tracking the Accuracy, Trustworthiness and Provenance of the Data

As we have pointed out in the introductory chapter, the exploration of Big Data involves checking the quality of the data and its trustworthiness. Recall that some data sources produce low quality or uncertain data, see e.g. tweets, blogs, and social media. Earlier lessons of mining real data sets have clearly showed that the accuracy of the results strongly depends on the quality of the data and the appropriateness of the pre-processing. Moreover, if the final models interact with the environment and/or are applied to critical domains of human activities, then a good verification of the input data and their pre-processing as well as the deployment of data mining results all become much more crucial than in earlier information systems.

Some of the authors of this book mention these issues in the context of the process of knowledge discovery see, e.g., chapter "Big Data and the Internet of Things" by M. Shah (in Sect. 9 where he presents his concerns about the limitations of current solutions for the Internet of Things). Moreover, we have briefly described the provenance challenges for Big Data chapter "A Machine Learning Perspective on Big Data Analysis", Sect. 2.

It is important to note that more efforts should be done and new innovative approaches are needed. Some authors argue that new methods are necessary due to the complexity of Big Data. We can refer the reader to such papers as [10, 11, 19] for more information on new methods considered in the context of Big Data provenance. The authors of [22] describe approaches that attempt to track the provenance of workflows for MapReduce jobs. Recording provenance in distributed environments is also considered in [32].

Provenance also opens up additional topics for machine learning research. For instance, in the case of dynamic and changing data, the evolutionary history and the origins of data items become more complicated. The authors of [7] claim that trust measures are not static and that learning approaches could be applied to discover new measures of interesting data sources using others sources. In particular, new unsupervised methods have been proposed in [52]. Other research [51] has also shown the usefulness of semi-supervised learning methods that start with a portion of ground truth data. It was also advocated in [7] that developing new innovative methods, which can run on parallel platforms and deal with scalable data and numerous heterogeneous sources is one of the highly desired future research directions in the field.

3.3 Data Visualization and Visual Data Mining

Data analysts use visualization tools to understand the unknown structure of data and underlying patterns. Many tools have been developed for multidimensional data or more structured data. The reader is referred to [20] for their review. These authors also describe several visual data mining tools that may facilitate interactive mining based on the user's judgment of intermediate data mining results. Some of them use special methods to visualize mining results, e.g. clustering or classifiers. Interaction mechanisms for filtering, querying, and selecting data are also available. However, it is claimed that such visual exploration is too often available as a separate tool while it should be more tightly coupled with analytical methods into one knowledge discovery system.

R. Sarmento, M. Oliveira, M. Cordeiro, and J. Gama discuss the practical usefulness of visualizing large telecommunication networks in chapter "Social Network Analysis in Streaming Call Graphs". To efficiently handle very large and dynamic graphs, the authors have to model them as a kind of data stream and use special sampling techniques.

However, one could notice that many visualization methods and software tools have been developed in the context of standard, static and smaller data sets and that they are limited when it comes to exploring big data sets. The scale and complexity of Big Data may be too critical a challenge for current techniques and their implementations.

Reports like [23] list other requirements to make new visualization systems suitable for Big Data. These are:

- Enabling real-time data analysis (computationally cost-effective),
- Using in-memory compression to enable the handling of large-scale data,
- Supporting the interactive exploration of the data at different stages and the fast presentation of reports,
- Showing meaningful results (e.g., with appropriate context information and special presentation techniques to overcome the difficulties associated with too many results),
- Allowing users to share their presentations and reports with others and to collaborate in a sufficient secure way.

Then, DeGeer in [12] noticed that traditional visualization tools are too oriented toward the presentation of what a user may already know about the data. Instead, they should be exploring unknown aspects - which is more characteristic for data mining or even previously for Exploratory Data Analysis in statistics [48]. Furthermore, DeGeer presents a postulate of what a stronger visual interactivity means: the user has to be able to explore the data "on the fly", change its interests, filter out irrelevant information, deal with outliers and isolate unexpected patterns. He also notices that existing visualization tools are good for static information but that they generally fail to work with dynamic data.

Real-time visualization is particularly useful in data streams. Systems should handle a large number of very fast updates and offer innovative ideas on how to present changes in the data structure. The authors of the comprehensive survey on the topic present a similar opinion [31]. They also give an example of an open problem concerning the quick detection of breaking news events from huge amounts of streaming tweets. Following more recent papers by data stream researchers [24], the visualization of concept drifts and the graphical evaluation of model reactions to them are still open problems. Moreover, Gaber et al. claim that there are currently no on-line real-time visualization tools to complement the Ubiquitous Data Stream Mining algorithms [17]. A final postulate is to construct efficient visualization-based data discovery tools for mobile devices.

Other research reports [23] show other opportunities for applying visualization techniques to the protection of data quality (helping to find errors [1]) and supporting tracks of data provenance (graphical display of user activity records, characteristics of data sources).

3.4 User Feedback Integration and Result Interpretation

Since the beginnings of knowledge discovery from data, it has been stressed that users/decision makers should be able understand the analysis and the results of the machine learning algorithms. These postulates are also valid for many Big Data applications. For instance, [40] describes the real world successful application of data mining to predict manhole explosions and fires in the New York electrical network. Black-box (non-transparent) predictive models were treated as neither useful

nor convincing. Every step of the process had to be verified by both scientists and company engineers. Therefore, the research team designed several software tools that allowed transparency of the main operations and provided reasons for the predictions made by the final system. This allowed the integration of domain expertise (by company specialists) into the modelling process, data verification, and system evaluation.

In [34], Stan Matwin pointed out that appropriate interpretation of the results may be more important than better accuracy of the models, in particular when results are used for making decisions concerning people, like medical diagnostics or administrative decisions. However, he also noted that a good interpretation is still a research challenge for the machine learning and data mining fields. A limited number of popular approaches mainly trees, rules, Bayesian networks—offer, so called, symbolic knowledge representations, which could be directly inspected and interpreted by humans. Measuring and evaluating the interpretation abilities offered by various learning algorithms is still less studied than other criteria. In his view, this question should be brought to the fore and treated in an inter-disciplinary manner. Visualization methods could partly support users in interpretation tasks.

Another issue is that, data sources may contain erroneous data, or applied algorithms may not meet all the assumptions and, as a result, may produce inaccurate results. Responsible users will not rely exclusively on computer calculations but, instead, will try to verify the results—which again should be supported by new developed techniques.

However, these expectations are real challenges for Big Data—due to data complexity, sophisticated workflow of data transformations, distributed processing, and the application of many algorithms. Similarly to studying data provenance, there is a need for capturing adequate metadata reports, and powerful visualization tools that could involve human experts into the analysis could help interpret analytical results.

This type of use for data mining systems calls for more adequate users' interaction facilities which would allow humans to provide feedback or guidance. Interactiveness has been relatively under-emphasized in the context of data mining [7]. However, it will become more important when dealing with Big Data properties, such as all "V" characteristics. For instance, user guidance can help narrow the massive data into reduced, promising sub-spaces and accelerate the processing. Users can also evaluate and interpret intermediate results, search for hypotheses directly, and repeat certain steps with different assumptions or parameters if necessary.

This means that beside designing good visualization tools, it is necessary to develop special infrastructures and carry out more advanced research on evaluation measures and validation procedures. In particular, this refers to situations where algorithms may produce too many results and where finding a limited number of interesting patterns is not an obvious task [2, 21].

4 Stan Matwin's Contributions to Big Data Analytics

Stan Matwin's contributions to Big Data Analytics are many and quite significant. They have impacted the field in many ways.

Although the issue was only briefly discussed in chapter "A Machine Learning Perspective on Big Data Analysis", the class imbalance problem has been and will remain a confounding problem for machine learning, data mining and Big Data Analysis for years to come. Matwin and his colleagues were some of the first researchers to address the issue in [25, 26]. The approach they proposed remains a popular way of solving the problem close to 20 years later. Their work also helped popularize the use of the geometric mean (G-Mean) in class imbalance problems [27]. This was important since, on the one hand, this measure is still used today and on the other hand, it was an early attempt to challenge the usefulness of accuracy as the sole criterion in all situations. This led to its gradual replacement by (or at least competition with) the AUC, Precision/Recall Curves, etc.

Another of Matwin's important contribution is in the area of Text Mining. As seen in Sect. 1 of this chapter, data will increasingly be coming from the Internet and, in particular, from Social Media. This means that text processing has been and will continue to be an extremely important area of research in Big Data Analysis. Matwin's most important contribution in this area has been in feature engineering—as discussed in Sect. 2.7 of this chapter[5]. Feature Engineering remains an important topic of research both in text mining and in biomedical applications—but he also contributed interesting results in the areas of co-training, name entity recognition, word sense recognition, etc. [30, 41, 42].

As discussed in Sect. 3.1 of this chapter, Matwin also became interested in the problem of Privacy in Data Mining long before it became a popular issue [54]. As early as 2002, he developed, together with students and colleagues, privacy-oriented Data Mining algorithms [14].

Matwin's interest in practical applications led him to work on a wide variety of problems, including predicting who in a hospital emergency room will need hospitalization, recognizing oil spills in the ocean, categorizing medical articles, detecting emerging trends in a political campaign or in public opinion. Overall, he has contributed to solving problems in such wide-ranging fields as neuro-ophthalmology, forestry, electronics, and many others.

In 2013, with this experience in hand, Matwin established the Institute for Big Data Analytics at Dalhousie University. The institute is thriving and currently includes 7 research professors (including 6, on the executive board), 3 postdoctoral fellows, 6 Ph.D. students and 8 M.Sc. students. Ongoing projects span the domains of global telecommunications services, home care, retirement living and nursing homes, Marine Ecology, Text, anesthetics and post-operative care, to name only a few. The Institute will also be hosting the prestigious Conference on Knowledge Discovery and Data Mining in 2017.

Acknowledgments The work of Jerzy Stefanowski was partially supported by the Polish National Science Center under Grant No. DEC-2013/11/B/ST6/00963. The work of Nathalie Japkowicz was

supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. ASA—Discovery with Data: Leveraging statistics and computer science to transform science and society. A report of a Working Group of the American Statistical Association (July 2, 2014)
2. Bayardo, R., Agrawal, R.: Mining the most interesting rules. In: Proceedings of the 5th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 145–154 (1999)
3. Borne, K.: Scientific data mining in astronomy. In: Next Generation of Data Mining, pp. 91–114. Taylor & Francis, CRC Press (2009)
4. Breiman, L.: Statistical modeling: the two cultures. Statistical Sciences, pp. 199–231 (2001)
5. Caropreso, M., Matwin, S., Sebastiani, F.: A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In: Text databases and document management: Theory and practice, pp. 78–102 (2001)
6. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: CRISP-DM 1.0 step-by-step data mining guide. Technical report, The CRISP-DM consortium (2000)
7. Che, D., Safran, M., Peng, Z.: From Big Data to Big Data mining: challenges, issues and opportunities. In: Hong B. et al. (eds) DASFAA Workshops, Springer, LNCS, vol. 7827, pp. 1–15, (2013)
8. Chen, M., Mao, S., Liu, Y.: Big data. A survey. Mob. New Appl. **19**, 171–209 (2014)
9. Crawford, K., Schultz, J.: Big data and due process: toward a framework to redress predictive privacy harms. Boston College Law Rev. **55**(1), 93–128 (2014), http://lawdigitalcommons.bc.edu/bclr/vol55/iss1/4
10. Dai, C., Lin, D., Bertino, E., Kantarcioglu, M.: An approach to evaluate data trustworthiness based on data provenance. In: Proceedings of the 5th VLDB Workshop on Secure Data Management, pp. 82–98 (2008)
11. Davidson, S., Freire, J.: Provenance and scientific workflows: challenges and opportunities. In: Proceedings of SIGMOD'08, (2008)
12. DeGeer, W.: What is Next in Big Data. Wired, 12 Feb (2014)
13. Dwork, C., Mulligan, D.: It is not privacy and it is not fair. Stanford Law Review, online 35, 3 Sept (2013)
14. Felty, A., Matwin, S.: Privacy-oriented data mining by proof checking. In: Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery—PKDD 2002, Springer LNAI, pp. 138–149, (2002)
15. Gaber, M., Stahl, F., Gomes, J.: Pocket Data Mining. Big Data on Small Devices. Series: Studies in Big Data (2014)
16. Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Rinzivillo, S., Trasarti, R.: Unveiling the complexity of human mobility by querying and mining massive trajectory data. VLDB J. **20**(5), 695–719 (2011)
17. Gillick, B., Gaber, M., Krishnaswamy, S., Zaslavsky, A.: Visualisation of cluster dynamics and change detection in ubiquitous data stream mining. Proc. IWUC'2006, 29–38 (2006)
18. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature **457**(7232), 1012–1014 (19 Feb 2009)
19. Glavic, B.: Big Data provenance: challenges and implications for benchmarking. In: Specifying Big Data Benchmarks, Springer, pp. 72–80, (2014)
20. Han, J., Gao, J.: Research challenges for data mining in science and engineering, In: Next Generation of Data Mining London: Chapman & Hall, pp. 1–18 (2009)
21. Hilderman, R.J., Hamilton, H.J.: Knowledge Discovery and Measures of Interest. Kluwer Academic, Boston (2002)

22. Ikeda, R., Park, H., Widom, J.: Provenance for generalized map and reduce workflows. In Proc. of CIDR, 273–283 (2011)
23. Intel White Paper: Big Data Visualization: Turning Big Data Into Big Insights—The Rise of Visualization-based Data Discovery Tools, (March 2013)
24. Krempl, G., Zliobaite, I., Brzezinski, D., Hullermeier, E., Last, M., Lemaire, V., Noack, T., Shaker, A., Sievi, S., Spiliopoulou, M., Stefanowski, J.: Open challenges for data stream mining research. ACM SIGKDD Explor. **16**(1), 1–10 (2014). June
25. Kubat, M., Holte, R., Matwin, S.: Machine learning for the detection of oil spills in satellite radar images. Mach. Learn. **30**(2–3), 195–215 (1998)
26. Kubat, M., Holte, R., Matwin, S.: Addressing the curse of imbalanced training sets: one-sided selection. Proc. ICML **97**, 179–186 (1997)
27. Kubat, M., Holte, R., Matwin, S.: Learning when negative examples abound. In: Proc. ECML '97, pp. 146–153 (1997)
28. Lally, A., et al.: Question analysis: how Watson reads a clue. IBM J. Res. Dev. **56**(3/4), (2012)
29. Lazer, D., Kennedy, R., King, G., Vespignani, A.: The parable of google flu: traps in big data analysis. Science, **343**, 1203–1205 (14 March 2014)
30. Li, X., Szpakowicz, S., Matwin, S.: A WordNet-based algorithm for word sense disambiguation. In Proc. IJCAI-95, pp. 1368–1374, (1995)
31. Liu, S., Cui, W., Wu, Y., Liu, M.: A survey on information visualization: recent advances and challenges. Vis. Comput. **30**(12), 1373–1393 (2014). December
32. Malik, T., Nistor, L., Gehani, A.: Tracking and sketching distributed data provenance. In: eScience, pp. 190–197 (2012)
33. Matwin's opinions on data privacy issues: http://www.dal.ca/faculty/computerscience/research-industry/researchchairs/stan_matwin.html (Retrieved 2015)
34. Matwin, S.: Machine learning: four lessons and what is next? Bull. Polish AI Soc. **2**, 2–7 (2013)
35. Matwin, S.: Privacy-preserving data mining techniques: survey and challenges. In Custers, B., Calders, T., Schermer, B., Zarsky T. (eds.) Discrimination and Privacy in the Information Society. Springer Series on Studies in Applied Philosophy, Epistemology and Rational Ethics, vol. 3, pp. 209–221 (2013)
36. Mayer-Schonberger, V., Cukier, K.: Big data: a revolution that will transform how we live, work and think. Eamon, Dolan/Houghton Mifflin Harcourt (2013)
37. Musolesi, M.: Big mobile data mining: good or evil? IEEE Internet Computing, pp. 2–5 (2014)
38. Pederschi, D., Calders, T., Custer, B.: Big Data mining, fairness and privacy a vision statement towards an interdisciplinary roadmap of research. KDnuggest Rev. **11**(26) (2011)
39. Richards, N., King, J.: Three paradoxes of big data. Stanford Law Rev. Online **66**, 41–46 (2013)
40. Rudin, C., Passonneau, R., Radeva, A., Jerome, S., Issac, D.: 21st century data miners meet 19-th century electrical cables. IEEE Computer, 103–105, (June 2011)
41. Scott, S., Matwin, S.: Text classification using WordNet hypernyms. In: Procedings of the Conference—Use of WordNet in Natural Language Processing Systems, pp. 38–44 (1998)
42. Scott, S., Matwin, S.: Feature engineering for text classification. Proc. ICML'99, 379–388 (1999)
43. Singh, D., Reddy, C.: A survey on platforms for big data analytics. J. Big Data **1**(8), 2–20 (2014)
44. Skowron, A., Stepaniuk, J., Swiniarski, R.: Modeling rough granular computing based on approximation spaces. Inf. Sci. **184**, 20–43 (2012)
45. Smailovic, J., Grcar, M., Lavrac, N., Znidarsic, M.: Stream-based active learning for sentiment analysis in the financial domain. Inf. Sci. **285**, 181–203 (2014)
46. Sun, Y., Han, J., Yan, X., Yu, P.: Mining knowledge from interconnected data: a heterogeneous information networks analysis approach. VLDB Endowment **5**(12), 2022–2023 (2012)
47. Teen, O., Polonetsky, J.: Privacy in the age of big data. A time for big decisions. Stanford Law Rev. Online **64**, 63–69 (2012)
48. Tukey, J.: Exploratory Data Analysis. Addison Wesley, Reading (1970)
49. Weisburd, D., Telep, C.: Hot spot policing: what we know and what we need to know. J. Contemp. Crim. Justice **30**, 200–220 (2014)

50. Working Paper on Big Data and Privacy—Privacy principles under pressure in the age of Big Data analytics—55th Meeting of International Working Group on Data Protection in Telecommunications, vol. 5, 6 May 2014, Skopje (2014)
51. Yin, X., Tan, W.: Semi-supervised truth discovery. In: Proceedings of the 20th International Conference on WWW, pp. 217–226 (2011)
52. Yin, X., Han, J., Yu, P.: Truth discovery with multiple conflicting information providers on the Web. In: Proceedings of the 13th ACM SIGKDD Conference on KDD, pp. 1048–1052 (2007)
53. Zhan, J., Chang, L., Matwin, S.: Privacy-preserving multi-party decision tree induction. In: Research Directions in Data and Applications Security, vol. XVIII, pp. 341–355 (2004)
54. Zhan, J., Matwin, S., Chang, L.: Privacy-preserving collaborative association rule mining. J. Netw. Comput. Appl. **30**(3), 1216–1227 (2007)