Studies in Big Data 16

Nathalie Japkowicz
Jerzy Stefanowski *Editors*

Big Data Analysis: New Algorithms for a New Society

Springer

Studies in Big Data

Volume 16

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

About this Series

The series "Studies in Big Data" (SBD) publishes new developments and advances in the various areas of Big Data-quickly and with a high quality. The intent is to cover the theory, research, development, and applications of Big Data, as embedded in the fields of engineering, computer science, physics, economics and life sciences. The books of the series refer to the analysis and understanding of large, complex, and/or distributed data sets generated from recent digital sources coming from sensors or other physical instruments as well as simulations, crowd sourcing, social networks or other internet transactions, such as emails or video click streams and other. The series contains monographs, lecture notes and edited volumes in Big Data spanning the areas of computational intelligence incl. neural networks, evolutionary computation, soft computing, fuzzy systems, as well as artificial intelligence, data mining, modern statistics and operations research, as well as self-organizing systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

More information about this series at http://www.springer.com/series/11970

Nathalie Japkowicz · Jerzy Stefanowski
Editors

Big Data Analysis: New Algorithms for a New Society

Springer

Editors
Nathalie Japkowicz
University of Ottawa
Ottawa, ON
Canada

Jerzy Stefanowski
Institute of Computing Sciences
Poznań University of Technology
Poznań
Poland

ISSN 2197-6503 ISSN 2197-6511 (electronic)
Studies in Big Data
ISBN 978-3-319-26987-0 ISBN 978-3-319-26989-4 (eBook)
DOI 10.1007/978-3-319-26989-4

Library of Congress Control Number: 2015955861

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

This book is dedicated to Stan Matwin in recognition of the numerous contributions he has made to the fields of machine learning, data mining, and big data analysis to date. With the opening of the Institute for Big Data Analytics at Dalhousie University, of which he is the founder and the current Director, we expect many more important contributions in the future.

Stan Matwin was born in Poland. He received his Master's degree in 1972 and his Ph.D. in 1977, both from the Faculty of Mathematics, Informatics and Mechanics at Warsaw University, Poland. From 1975 to 1979, he worked in the Institute of Computer Science at that Faculty as an Assistant Professor. Upon immigrating to Canada in 1979, he held a number of lecturing positions at Canadian universities, including the University of Guelph, York University, and Acadia University. In 1981, he joined the Department of Computer Science (now part of the School of Electrical Engineering and Computer Science) at the University of Ottawa, where he carved out a name for the department in the field of machine learning over his 30+ year career there (he became a Full Professor in 1992, and a Distinguished University Professor in 2011). He simultaneously received the State Professorship from the Republic of Poland in 2012.

He founded the Text Analysis and Machine Learning (TAMALE) lab at the University of Ottawa, which he led until 2013. In 2004, he also started cooperating as a "foreign" professor with the Institute of Computer Science, Polish Academy of Sciences (IPI PAN) in Warsaw. Furthermore, he was invited as a visiting researcher or professor in many other universities in Canada, USA, Europe, and Latin America, where in 1997 he received the UNESCO Distinguished Chair in Science and Sustainable Development (Universidad de Sao Paulo, ICMSC, Brazil).

In addition to his position as professor and researcher, he served in a number of organizational capacities: former president of the Canadian Society for the Computational Studies of Intelligence (CSCSI), now the Canadian Artificial Intelligence Society (CAIAC), and of the IFIP Working Group 12.2 (Machine Learning), Founding Director of the Information Technology Cluster of the Ontario Research Centre for Electronic Commerce, Chair of the NSERC Grant Selection

Committee for Computer Science, and member of the Board of Directors of Communications and Information Technology Ontario (CITO).

Stan Matwin is the 2010 recipient of the Distinguished Service Award of the Canadian Artificial Intelligence Society (CAIAC). He is Fellow of the European Coordinating Committee for Artificial Intelligence and Fellow of the Canadian Artificial Intelligence Society.

His research spans the fields of machine learning, data mining, big data analysis and their applications, natural language processing and text mining, as well as technological aspects of e-commerce. He is the author and co-author of over 250 research papers.

In 2013, he received the Canada Research Chair (Tier 1) in Visual Text Analytics. This prestigious distinction and a special program funded by the federal government allowed him to establish a new research initiative. He moved to Dalhousie University in Halifax, Canada, where he founded, and now directs, the Institute for Big Data Analytics.

The principal aim of this Institute is to become an international hub of excellence in Big Data research. Its second goal is to be relevant to local industries in Nova Scotia, and in Canada (with respect to applications relating to marine biology, fisheries and shipping). Its third goal is to develop a focused and advanced training program that covers all aspects of big data, preparing the next generation of researchers and practitioners for research in this field of study.

On the web page of his Institute, he presents his vision on Big Data Analytics. He stresses, “Big data is not a single breakthrough invention, but rather a coming together and maturing of several technologies: huge, inexpensive data harvesting tools and databases, efficient, fast data analytics and data mining algorithms, the proliferation of user-friendly data visualization methods and the availability of affordable, massive and non-proprietary computing. Using these technologies in a knowledgeable way allows us to turn masses of data that get created daily by businesses and the government into a big asset that will result in better, more informed decisions.”

He also recognizes the potential transformative role of big data analysis, in that it could support new solutions for many social and economic issues in health, cities, the environment, oceans, education access, personalized medicine, etc. These opinions are reflected in the speech he gave at the launch of his institute, where his recurring theme was “Make life better.” His idea is to use big data (i.e., large and constantly growing data collections) to learn how to do things better. For example, he proposes to turn data into an asset by, for instance, improving motorized traffic in a big city or ship traffic in a big port, creating personalized medical treatments based on a patient's genome and medical history, and so on.

Notwithstanding the advantages of big data, he also recognizes its risks for society, especially in the area of privacy. As a result, since 2002, he has been engaged in research on privacy preserving data mining.

Other promising research directions, in his opinion, include data stream mining, the development of new data access methods that incorporate sharing ownership mechanisms, and data fusion (e.g., geospatial applications).

We believe that this book reflects Stan Matwin's call for careful research on both the opportunities and the risks of Big Data Analytics, as well as its impact on society.

Nathalie Japkowicz
Jerzy Stefanowski

Acknowledgments

We take this opportunity to thank all contributors for submitting their papers to this edited book. Their joint efforts and good co-operation with us have enabled to successfully finalize the project of this volume.

Moreover, we wish to express our gratitude to the following colleagues who helped us in the reviewing process: Anna Kobusińska, Ewa Łukasik, Krzysztof Dembczyński, Miłosz Kadziński, Wojciech Kotłowski, Robert Susmaga, Andrzej Szwabe on the Polish side and Vincent Barnabe-Lortie, Colin Bellinger, Norrin Ripsman and Shiven Sharma on the Canadian side.

Continuous guidance and support of the Springer Executive Editor Dr. Thomas Ditzinger and Springer team are also appreciated. Finally, we owe a vote of thanks to Professor Janusz Kacprzyk who has invited us to start the project of this book and has supported for our efforts.

Contents

A Machine Learning Perspective on Big Data Analysis

Nathalie Japkowicz and Jerzy Stefanowski

Abstract This chapter surveys the field of Big Data analysis from a machine learning perspective. In particular, it contrasts Big Data analysis with data mining, which is based on machine learning, reviews its achievements and discusses its impact on science and society. The chapter concludes with a summary of the book's contributing chapters divided into problem-centric and domain-centric essays.

1 Preliminaries

In 2013, Stan Matwin opened the Institute for Big Data Analytics at Dalhousie University. The Institute's mission statement, posted on the website is, "To create knowledge and expertise in the field of Big Data Analytics by facilitating fundamental, interdisciplinary and collaborative research, advanced applications, advanced training and partnerships with industry." In another position paper [46] he posited that Big Data sets the new problems they come with and represent the challenges that machine learning research needs to adapt to. In his opinion, Big Data Analytics will significantly influence the field with respect to developing new algorithms as well as in the creation of applications with greater societal importance.

The purpose of this edited volume, dedicated to Stan Matwin, is to explore, through a number of specific examples, how the study of Big Data analysis, of which his institute is at the forefront, is evolving and how it has started and will most likely continue to affect society. In particular, this book focuses on newly developed algorithms affecting such areas as business, financial forecasting, human mobility, the Internet of Things, information networks, bioinformatics, medical systems and life science.

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

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

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_1

Moreover, this book will provide methodological discussions about the principles of mining Big Data and the difference between traditional statistical data analysis and newer computing frameworks for processing Big Data.

This chapter is divided into three sections. In Sect. 2, we define Big Data Analysis and contrast it with traditional data analysis. In Sect. 3, we discuss visions about the changes in science and society that Big Data brings about, along with all of their benefits. This is countered by warnings about the negative effects of Big Data Analysis along with its pitfalls and challenges. Section 4 introduces the work that will be presented in the subsequent chapters and fits it into the framework laid out in Sects. 2 and 3. Conclusions about the research presented in this book will be presented in the final chapter along with a review of Stan Matwin's contributions to the field.

2 What Do We Call Big Data Analysis?

For a traditional Machine Learning expert, "Big Data Analysis" can be both exciting and threatening. It is threatening in that it makes a lot of the research done in the past obsolete since previously designed algorithms may not scale up to the amount of new data now typically processed, or they may not address the new problems generated by Big Data Analysis. In addition, Big Data analysis requires a different set of computing skills from those used in traditional research. On the other hand, Big Data analysis is exhilarating because it brings about a multitude of new issues, some already known, and some still to be discovered. Since these new issues will need to be solved, Big Data analysis is bringing a new dynamism to the fields of Data Mining and Machine Learning.

Yet, what is Big Data Analysis, really? In this section and this introductory chapter, in general, we try to figure out what Big Data Analysis really is, at least as far as Machine Learning scientists are concerned, whether it is truly different from what Machine Learning scientists have been doing in the past, and whether it has the potential, heralded by many, to change society in a dramatic way or whether the changes will be incremental and relatively small. Basically, we are trying to figure out what all the excitement is about! We begin by surveying some general definitions of mining Big Data, discussing characteristic features of these data and then move on to more specific Machine Learning issues. After discussing a few well-known successful applications of Big Data Analysis, we conclude this section with a survey of the specific innovations in Machine Learning and Data Mining research that have been driven by Big Data Analysis.

2.1 General Definitions of Big Data

The original and much cited definition from the Gartner project [10] mentions the "three *V*s": Volume, Velocity and Variety. These "*V* characteristics" are usually explained as follows:

Volume—the huge and continuously increasing size of the collected and analyzed data is the first aspect that comes to mind. It is stressed that the magnitude of Big Data is much larger than that of data managed in traditional storage systems. People talk about terabytes and petabytes rather than gigabytes. However, as noticed by [35], the size of Big Data is "a constantly moving target- what is considered to be Big today will not be so years ahead".

Velocity—this term refers to the speed at which the data is generated and input into the analyzing system. It also forces algorithms to process data and produce results in limited time as well as with limited computer resources.

Variety—this aspect indicates heterogeneous, complex data representations. Quite often analysts have to deal with structured as well as semi-structured and unstructured data repositories.

IBM added a fourth "*V*" which stands for "**Veracity**". It refers to the quality of the data and its trustworthiness. Recall that some sources produce low quality or uncertain data, see e.g. tweets, blogs, social media. The accuracy of the data analysis strongly depends on the quality of the data and its pre-processing.

In 2011, IDC added, yet, another dimension to Big Data analysis: "**Value**". Value means that Big Data Analysis seeks to economically extract value from very large volumes of a wide variety of data [16]. In other words, mining Big Data should provide novel insights into data, application problems, and create new economical value that would support better decision making; see some examples in [22].

Another "*V*", still, is for "**Variability**". Authors of [23] stress that there are changes in the structure of the data, e.g. inconsistencies which can be shown in the data as time goes on, as well as changes in how users want to interpret that data.

These are only a few of the definitions that have previously been proposed for Big Data Analysis. For an excellent survey on the topic, please see [20].

Note, however, that most of these definitions are general and geared at business. They are not that useful for Machine Learning Scientists. In this volume, we are more interested in a view of Big Data as it relates to Machine Learning and Data Mining research, which is why we explore the meaning of Big Data Analysis in that context next.

2.2 Machine Learning and Data Mining Versus Big Data Analysis

One should remember that data mining or more generally speaking the field of Knowledge Discovery from Databases started in the late 1980s [50]—before the appearance of Big Data applications and research. Machine Learning is an older

research discipline that provided many algorithms for carrying out data mining steps or inspired more specialized and complex solutions. From a methodological point of view, it strongly intersects with the field of data mining. Some researchers even identify traditional data mining with Machine Learning while others indicate differences, see e.g., discussions in [33, 41, 42].

It is not clear that there exists a simple, clear and concise Big Data definition that applies to Machine Learning. Instead, what Machine Learning researchers have done is list the kinds of problems that may arise with the emergence of Big Data. We now present some of these problems in the table below (due to its length it is actually split into two Tables 1 and 2). The table is based on discussions in [23, 24], but we organized them by categories of problems. We also extended these categories according to our own understanding of the field. Please note that some of the novel aspects of Big Data characteristics have already been discussed in the previous subsection, so here, we only mention those that relate to machine learning approaches in data mining.

Please note, as well, that this table is only an approximation: as with the boundary between machine learning and data mining, the boundary between the traditional data mining discipline and the Big Data analysis discipline is not clear cut. Some issues listed in the Big Data Analysis column occurred early on in the discipline, which could still be called traditional data mining. Similarly, some of the issues listed in the Data Mining column may, in fact, belong more wholly to the Big Data Analysis column even if early, isolated work on these problems had already started before the advent of the Big Data Analysis field. This is true at the data set level too: the distinction between a Data Mining problem and a Big Data Analysis problem is not straightforward. A problem may encounter some of the issues described in the "Big Data Analysis" column and still qualify as a data mining problem. Similarly, a problem that qualifies as a "Big Data" problem may not encounter all the issues listed in the "Big Data Analysis" column.

Once again, this table serves as an indication of what the term "Big Data Analysis" refers to in machine learning/data mining. The difference between traditional data mining and Big Data Analysis and most particularly, the novel elements introduced by Big Data Analysis, will be further explored in Sect. 2.4 where we look at specific problems that can and must now be considered. Prior to that, however, we take a brief look at a few successful applications of Big Data Analysis in the next section.

2.3 Some Well-Known and Successful Applications of Big Data Analysis

Big Data analysis has been successfully applied in many domains. We limit ourselves to listing a few well known applications, though these applications and many others are quite interesting and would have been given greater coverage if space restrictions had not been a concern:

Table 1 Part A—Traditional data mining versus big data analysis with respect to different aspects of the learning process

	Traditional data mining	Big data analysis
Memory access	The data is stored in centralized RAM and can be efficiently scanned several times	The data may be stored on highly distributed data sources In case of huge, continuous data streams, data is accessed only in a single scan and limited subsets of data items are stored in memory
Computational processing and architectures	Serial, centralized processing is sufficient A single-computer platform that scales with better hardware is sufficient	Parallel and distributed architectures may be necessary Cluster platforms that scale with several nodes may be necessary
Data types	The data source is relatively homogeneous The data is static and, usually, of reasonable size	The data may come from multiple data sources which may be heterogeneous and complex The data may be dynamic and evolving. Adapting to data changes may be necessary.
Data management	The data format is simple and fits in a relational database or data warehouses. Data management is usually well-structured and organized in a manner that makes search efficient. The data access time is not critical	Data formats are usually diverse and may not fit in a relational database The data may be greatly interconnected and needs to be integrated from several nodes Often special data systems are required that manage varied data formats (NoSQL databases, Hadoop or Spark platforms, etc.) The data access time is critical for scalability and speed
Data quality	The provenance and pre-processing steps are relatively well documented Strong correction techniques were applied for correcting data imperfection Sampling biases can, somehow, be traced back The data is relatively well tagged and labeled	The provenance and pre-processing steps may be unclear and undocumented There is a large amount of uncertainty and imprecision in the data Sampling biases are unclear Only a small number of data are tagged and labeled

Table 2 Part B—Traditional data mining versus big data analysis with respect to different aspects of the learning process

	Traditional data mining	Big data analysis
Data handling	Security and Privacy are not of great concern Policies about data sharing are not necessary	Security and Privacy may matter Data may need to be shared and the sharing must be done appropriately
Data processing	Only batch learning is necessary Learning can be slow and off-line The data fits into memory All the data has some sort of utility The curse of dimensionality is manageable No compression and minimal sampling is necessary Lack of sufficient data is a problem	Data may arrive in a stream and need to be processed continuously Learning may need to be fast and online The scalability of algorithms is important The data may not fit in memory The useful data may be buried in a mass of useless data The curse of dimensionality is disproportionate Compression or sampling techniques must be applied Lack of sufficient data of interest remains a problem
Result analysis and integration	Statistical significance results are meaningful Many visualization tools have been developed Interaction with users is well developed The results do not usually need to be integrated with other components	With massive data sets, non-statistically significant results may appear statistically significant Traditional visualization software may not work well with massive data The results of the Big Data analysis may need to be integrated with other components

- Google Flu Trends—Researchers at Google and the Center for Disease Control (CDC) teamed together to build and analyse a surveillance system for early detection of flu epidemics, which is based on tracking a different kind of information from flu-related web search queries [29].[1]
- Predicting the Next Deadly Manhole Explosion in New York Electric Network—In 2004 the Con Edison Company began a proactive inspection program, with the goal of finding the places in New York's network of electrical cables where trouble was most likely to strike. The company co-operated with a research team at

[1] While this application was originally considered a success, it subsequently obtained disappointing results and is now in the process of getting improved [4].

Columbia University to develop an algorithm that predicts future manhole failure and could support the company's inspection and repair programs [55, 56].

- Wal-Mart's use of Big Data Analytics—Wal-Mart has been using Big Data Analysis extensively to achieve a more efficient and flexible pricing strategy, better-managed advertisement campaigns and a better management of their inventory [36].
- IBM Watson—Watson is the famous Q&A Computer System that was able to defeat two former winners of the TV game show *Jeopardy*! in 2011 and win the first prize of one million dollars. This experiment shows how the use of large amounts of computing power can help clear up bottlenecks when constructing a sophisticated language understanding module coupled with an efficient question answering system [63].
- Sloan Sky Digital Survey—The Sloan Sky Digital Survey has gathered an extensive collection of images covering more than a quarter of the sky. It also created three-dimensional maps of over 930,000 galaxies and 120,000 quasars. This data is continuously analyzed using Big Data Analytics to investigate the origins of the universe [60].
- FAST—Homeland Security FAST (Future Attribute Screening Technology) intends to detect whether a person is about to commit a crime by monitoring the contractions of their facial muscles, which are, believed to reflect seven primary emotions and emotional cues linked to hostile intentions. Such a system would be deployed at airports, border crossings and at the gates of special events, but it is also the subject of controversy due to its potential violation of privacy and the fact that it would probably yield a large number of false positives [25].

The reviews of additional applications where Big Data Analysis has already proven itself worthy can be found in other articles such as [15, 16, 67].

2.4 Machine Learning Innovations Driven by Big Data Analysis

In this subsection, we look at the specific new problems that have emanated from the handling of Big Data sets, and the type of issues they carry with them. This is an expansion of the information summarized in Tables 1 and 2.

We divide the problems into two categories:

1. The completely new problems which were never considered, or considered only in a limited range, prior to the advent of Big Data analysis, and originate from the format in which data is bound to present itself in Big Data problems.
2. The problems that already existed but have been disproportionately exaggerated since the advent of Big Data analysis.

2.4.1 New Problems Caused by the Format in Which Big Data Presents Itself

We now briefly discuss problems that were, perhaps, on the mind of some researchers prior to the advent of Big Data, but that either did not need their immediate attention since the kind of data they consider did not seem likely to occur immediately, or were already tackled but need new considerations given the added properties of Big Data sets. The advent and fast development of the Internet and the capability to store huge volumes of data and to process them quickly changed all of that. New forms of data have emerged and are here to stay, requiring new techniques to deal with them. These new kinds of data and the problems associated with them are now presented.

Graph Mining Graphs are ubiquitous and can be used to represent several kinds of networks such as the World Wide Web, citation graphs, computer networks, mobile and telecommunication networks, road traffic networks, pipeline networks, electrical power-grids, biological networks, social networks and so on. The purpose of graph mining is to discover patterns and anomalies in graphs and use these discoveries to useful means such as fraud detection, cyber-security, or social network mining (which will be discussed in more detail below). There are two different types of graph analyses that one may want to perform [37]:

- Structure analysis (which allows the discovery of patterns and anomalies in connected components, the monitoring of the radius and diameter of graphs, their structure, and their evolution).
- Spectral Analysis (which allows the analysis of more specific information such as tightly connected communities and anomalous nodes).[2]

Structure analysis can be useful for discovering anomalously connected components that may signal anomalous activity; it can also be used to give us an idea about the structure of graphs and their evolution. For example, [37] discovered that large real-world graphs are often composed of a "core" with small radius, "whiskers" that still belong to the core but are more loosely connected to it and display a large radius, and finally "outsiders" which correspond to disconnected components each with a small radius.

Spectral Analysis, on the other hand, allows for much more pinpointed discoveries. The authors of [37] were able to find that adult content providers create many twitter accounts and make them follow each other so as to look more popular. Spectral Analysis can thus be used to identify specific anomalous (and potentially harmful) behaviour that can, subsequently, be eliminated.

Mining Social Networks The notion of social networks and social network analysis is an old concept that emanated from Sociology and Anthropology in the 1930s [7]. The earlier research was focused on analysing sociological aspects of personal relationships using rigorous data collection and statistical analysis procedures. However,

[2] Please note that graphs were sometimes considered in traditional data mining (e.g., as structures of chemical compounds), but the graphs in question were of much smaller size than those considered today.

with the advent of social media, this kind of study recently took a much more concrete turn since all traces of social interactions through these networks are tangible.

The idea of Social Network Analysis is that by studying people's interactions, one can discover group dynamics that can be interesting from a sociological point of view and can be turned into practical uses. That is the purpose of mining social networks which can, for example, be used to understand people's opinions, detect groups of people with similar interests or who are likely to act in similar ways, determine influential people within a group and detect changes in group dynamics over time [7].

Social Network Mining tasks include:

- Group detection (who belongs to the same group?),
- Group profiling (what is the group about?),
- Group evolution (understanding how group values change),
- Link prediction (predict when a new relationship will form).

Social Network Analysis Applications are of particular interest in the field of business since they can help products get advertised to selected groups of people likely to be interested, can encourage friends to recommend products to each other, and so on.

Dealing with Different and Heterogeneous Data Sources Traditional machine learning algorithms are typically applied to homogeneous data sets, which have been carefully prepared or pre-processed in the first steps of the knowledge discovery process [50]. However, Big Data involves many highly heterogeneous sources with data in different formats. Furthermore, these data may be affected by imprecision, uncertainty or errors and should be properly handled. While dealing with different and heterogeneous data sources, two issues are at stake:

1. How are similar data integrated to be presented in the same format prior to analysis?
2. How are data from heterogeneous sources considered simultaneously in the analysis?

The first question belongs primarily to the area of designing data warehouses. It is also known as the problem of data integration. It consists of creating a unified database model containing the data from all the different sources involved.

The second question is more central to data mining as it may lead researchers to abandon the construction of a single model from integrated and transformed data in favour of an approach that builds several models from homogeneous subsets of the overall data set and integrates the results together (see e.g., [66]).

Combining the questions on graph mining discussed in the previous subsections and on heterogeneous sources discussed here leads to a commonly encountered problem: that of analysing a network of heterogeneous data (e.g., the nodes of the network represent people, documents, photos, etc.). This started a new sub-field called heterogeneous information network analysis [61], which consists of using a network scheme listing meta-information about the nodes and the links.

Data Stream Mining In the past, most of machine learning research and applications were focused on batch learning from static data sets. These, usually not massive, data sets were efficiently stored in databases or file systems and, if needed, could be accessed by algorithms multiple times. Moreover, the target concepts to be learned were well defined and stable. In some recent applications, learning algorithms have had to act in dynamic environments, where data are continuously produced at high speed. Examples of such applications include sensor networks, process monitoring, traffic management, GPS localizations, mobile and telecommunication call networks, financial or stock systems, user behaviour records, or web log analysis [27]. In these applications, incoming data form a data stream characterized by a huge volume of instances and a rapid arrival-rate which often requires a quick, real-time response.

Data stream mining, therefore, assumes that training examples arrive incrementally one at a time (or in blocks) and in an order over which the learning algorithm has no control. The learning system resulting from the processing of that data must be ready to be applied at any time between the arrivals of two examples, or consecutive portions (blocks) of examples [11]. Some earlier learning algorithms, like Artificial Neural Networks or Naive Bayes, were naturally incremental. However, the processing of data streams imposes new computational constraints for algorithms with respect to memory usage, limited learning and testing time, and single scanning of incoming instances [21]. In practice, incoming examples can be inspected briefly, cannot all be stored in memory, and must be processed and discarded immediately in order to make room for new incoming examples. This kind of processing is quite different from previous data mining paradigms and has new implications on constructing systems for analysing data streams.

Furthermore, with stream mining comes an important and not insignificant challenge: these algorithms often need to be deployed in dynamic, non-stationary environments where the data and target concepts change over time. These changes are known as concept drifts and are serious obstacles to the construction of a useful stream-mining system [39].

Finally, from a practical point of view, mining data streams is an exciting area of research as it will lead to the deployment of ubiquitous computing and smart devices [26].

Unstructured or Semi-Structured Data Mining Most Big Data sets are not highly structured in a way that can be stored and managed in relational databases. According to many reports, the majority of collected data sets are semi-structured, like in the case of data in HTML, XML, JSON or bibtex format, or unstructured, like in the case of text documents, social media forums or sound, images or video format [1]. The lack of a well-defined organization for these data types may lead to ambiguities and other interpretation problems for standard data mining tools.

The typical way to deal with unstructured data sets is to find ways to impose some structure on them and/or transform them into another representation, in order to be able to process them with existing data mining tools. In text mining, for example, it is customary to find a representation of the text using Natural Language Processing and Text Analytic tools. These include tools for removing redundancies and inconsistencies, tokenization, eliminating stop words, stemming, identification of terms

based on unigrams, bigrams, phrases or other features of the text which could lead to vector space models [43]. Some of these methods may also require collecting reference corpora of documents. Similar approaches are used for images or sound where high-level features are defined and used to describe the data. These features can then be processed by traditional learning systems.

Spatio-Temporal Data Mining Spatio-temporal data corresponds to data that has both temporal and spatial characteristics. The temporal characteristics refer to the fact that over time certain changes apply to the object under consideration and these changes are recorded at certain time intervals. The spatial aspect of the data refers to the location and shape of the object. Typical spatio-temporal applications include environment and climate (global change, land-use classification monitoring), the evolution of an earthquake or a storm over time, Public Health (monitoring and predicting the spread of disease), public security (finding hotspots of crime), geographical maps and census analysis, geo-sensor measurement networks, transportation (traffic monitoring, control, traffic planning, vehicle navigation), tracking GPS/mobile and localization-based services [54, 57, 58].

Handling spatio-temporal data is particularly challenging for different reasons. First, these data sets are embedded in continuous spaces, whereas typical data are often static and discrete. Second, classical data mining tends to focus on discovering the global patterns of models while in spatio-temporal data mining there is more interest on local patterns. Finally, spatio-temporal processing also includes aspects that are not present with other kinds of data processing. For example, geometric and temporal computations need to be included in the processing of the data, normally implicit spatial and temporal relationships need to be explicitly extracted, scale and granularity effects in space and time need to be considered, the interaction between neighbouring events has to be considered, and so on [65]. Moreover, the standard assumption regarding sample independence is generally false because spatio-temporal data tends to be highly correlated.

Issues of Trust/Provenance Early on, data mining systems and algorithms were typically applied to carefully pre-processed data, which came from relatively accurate and well-defined sources, thus trust was not a critical issue. With emerging Big Data, the data sources have many different origins, which may be less known and not all verifiable [15]. Therefore, it is important to be aware of the provenance of the data and establish whether or not it can be trusted [17]. Provenance refers to the path that the data has followed before arriving at its destination and Trust refers to whether both the source and the intermediate nodes through which the database passed are trustworthy.

Typically, data provenance explains the creation process and origins of the data records as well as the data transformations. Note that provenance may also refer to the type of transformation [58] that the data has gone through, which is important for people analysing it afterwards (in terms of biases in the data). Additional metadata, such as conditions of the execution environment (the details of software or computational system parameters), are also considered as provenance.

Data provenance has previously been studied in the database, workflow and geographical information systems communities [18]. However, the world of Big Data is

much more challenging and still not sufficiently explored. The main challenges in Big Data Provenance come from working with:

- massive scales of sources and their inter-connections as well as highly unstructured and heterogeneous data (in particular, if users also apply ad-hoc analytics, then it is extremely difficult to model provenance [30]);
- complex computational platforms (if jobs are distributed onto many machines, then debugging the Big Data processing pipeline becomes extremely difficult because of the nature of such systems);
- data items that may be transformed several times with different analytical pieces of software;
- extremely long runtimes (even with more advanced computational systems, analysing provenance and tracking errors back to their sources may require unacceptably long runtimes);
- difficulties in providing sufficiently simple and transparent programming models as well as high dynamism and evolution of the studied data items.

It is therefore an issue to which consideration must be given, especially if we expect the systems resulting from the analysis to be involved in critical decision making.

Privacy Issues Privacy Preserving Data Mining deals with the issue of performing data mining, i.e., drawing conclusions about the entire population, while protecting the privacy of the individuals on whose information the processing is done. This imposes constraints on the regular task of data mining. In particular, ways have to be found to mask the actual data while preserving its aggregate characteristics. The result of the data mining process on this constrained data set needs to be as accurate as if the constraint were not present [45].

Although privacy issues had been noticed earlier, they have become extremely important with the emergence of mining Big Data, as the process often requires more personal information in order to produce relevant results. Instances of systems requiring private information include localization-based and personalized recommendations or services, targeted and individualized advertisements, and so on. Systems that require a user to share his geo-location with the service provider are of particular concern since even if the user tries to hide his personal identity, without hiding his location, his precautions may be insufficient—the analysts could infer a missing identity by querying other location information sources. Barabasi et al. have, indeed, shown that there is a close correlation between people's identities and their movement patterns [31].

In social data sets, the privacy issue is particularly problematic since such sets usually contain many highly interconnected pieces of personal information. Even if the basic records could, somehow, be blocked from public view, a lot of personal information can be found and mined out when links to other data are found. At this point, all the pieces of information about a given person will be integrated and privacy compromised. Cukier and Mayer-Schoenberger describe several such case studies in their book [47]; see, for example, the surprising results obtained by an experimental analysis of old queries provided by AOL. Although the personal names

and IP were anonymized, researchers were able to correctly identify a single person by looking at associations between particular search phrases and additional data [6]. A similar situation occurred in the Netflix Prize Datasets, where researchers discovered correlations of ranks similar to those found in data sets from other services that used the users' full names [49]. This allowed them to clearly identify the anonymized users of the Netflix data.

This concludes our review of new problems that stemmed from the emergence of Big Data sets. We now move to existing problems that were amplified by the advent of Big Data.

2.4.2 Existing Problems Disproportionately Exaggerated by Big Data

Although the learning algorithms derived in the past were originally developed for relatively small data sets, it is worth noting that machine learning researchers have always been aware of the computational efficiency of their algorithms and of the need to avoid data size restrictions. Nonetheless, these efforts are not sufficient to deal with the flood of data that Big Data Analysis brought about. The two main problems with Big Data analysis, other than the emergence of new data format as discussed in previous subsections, consequently, are that:

1. The data is too big to fit into memory and is not sufficiently managed by typical analytical systems using databases.
2. The data is not, currently, processed efficiently enough.

The first problem is addressed by the design of distributed platforms to store the data and the second, by the parallelization of existing algorithms [15]. Some efforts have already been made in both directions and these are, now, briefly presented.

Distributed Platforms and Parallel Processing There have been several ventures aimed at creating distributed processing architectures. The best known one, currently, is the pioneering one introduced by Google. In particular, Google created a programming model called MapReduce which works hand in hand with a distributed file system called Google File System (GFS). Briefly speaking, MapReduce is a framework for processing parallelizable problems over massive data sets using a large number of computer nodes that construct a computational cluster. The programming consists of two steps: map and reduce. At the general level, map procedures read data from the distributed file system, process them locally and generate intermediate results, which are aggregated by reduce procedures into a final output. The framework also provides the distributed shuffle operations (which manage communication and data transfers), the orchestration of running parallel tasks, and deals with redundancy and fault tolerance.

Yahoo and other companies emulated the MapReduce architecture in an open-source framework. That Apache version of MapReduce is called Hadoop MapReduce and uses the Hadoop Distributed File System (HDFS), which is the open-source Apache equivalent of GFS [32]. The term Hadoop also refers to the collection of

additional software wrappers that can be installed on top of Hadoop and MapReduce, and can provide programmers with a better environment, see, for example, Apache Pig (SQL-like environment), Apache Hive (Hive is a warehouse system that conquers and analyses files stored in HDFS) and Apache HBase (a massive scale database management system) [59].

Hadoop and MapReduce are not the only platforms around. In fact, they have several limitations: most importantly, MapReduce is inefficient for running iterative algorithms, which are often applied in data mining. A few new fresh platforms have recently been developed to deal with this issue. The Berkeley Data Analytics Stack (BDAS) [9] is the next generation open-source data analysis tool for computing and analysing complex data. In particular, the BDAS component, called Spark, represents a new paradigm for processing Big Data, which is an alternative to Hadoop and should overcome some of its I/O limitations and eliminate some disk overhead in running iterative algorithms. It is reported that for some tasks it is much faster than Hadoop. Several researchers claim that Spark is better designed for processing machine learning algorithms and has much better programming interfaces. There are also several Spark wrappers such as Spark Streaming (large scale real time stream processing), GraphX (distributed graph system), and MLBase/Mlib (distributed machine learning library based on Spark) [38]. Other competitive platforms are ASTERIX or SciDB. Furthermore, specialized platforms for processing data streams include Apache S4 and Storm.

The survey paper [59] discusses criteria for evaluating different platforms and compares their application dependent characteristics.

Parallelization of Existing Algorithms In addition to the Big Data platforms that have been developed by various companies and, in some cases, made available to the public through open source platforms, a number of machine learning algorithms have been parallelized and placed in software packages made available to the public through open source channels.

Here is a list of some of the most popular open source packages:

- Apache's Mahout [40] which includes many implementations of distributed or otherwise scalable machine learning algorithms focused primarily on the areas of collaborative filtering, clustering and classification. Many of the implementations originally used the Apache Hadoop and MapReduce framework. However, some researchers judged that the implementations are too slow and the package not user-friendly [15]. In April 2014 the Mahout community decided to move its codebase onto newer data processing systems, such as Apache Spark, that offer a richer programming model and more efficient execution than Hadoop and MapReduce.
- BC-PDM (Big Cloud-Parallel Data Mining) is a cloud based series of implementations also based on Hadoop. It also supports parallel ETL (Extraction Transformation Load) processes and is more applicable to industrial Business Intelligence.
- MOA is an open source software package for stream data mining and contains implementations of classifiers, regression, clustering and frequent set mining [11]. Another newer, related project for distributed stream mining is the SAMOA project [48].

- NIBLE is yet another portable toolkit for implementing parallel ML-DM algorithms and runs on top of Hadoop [28].
- VowpalWabbit was developed by Yahoo and Microsoft Research. Its main aim is to provide efficient scalable implementations of online machine learning and support for a number of machine learning reductions, importance weighting, and a selection of different loss functions and optimization algorithms.
- h2o is the most recent open source mathematical and machine learning software for Big Data, released by Oxdata in 2014 [62]. It offers distribution and parallelism to powerful algorithms and allows programmers to use the R and JSON languages as APIs. It can be run on the top of either Hadoop or Spark.
- Graph mining tools are often used in mining Big Data. PEGASUS (Peta-scale Graph Mining System) is an open source package specifically designed for graph mining and also based on Hadoop. Giraph and GraphLab are two other such systems for Graph Mining.

A comprehensive survey of the various efforts made to scale up algorithms for parallel and distributed platforms can be found in the book entitled "Scaling Up Machine Learning. Parallel and Distributed Approaches" [8].

This concludes our general overview of Big Data Analysis from a Machine Learning point of view. The next section will discuss the scientific and societal changes that Big Data Analysis has led to.

3 Is Big Data Analysis a Game Changer?

In this section, we discuss the visions of people who believe that the foundations of science and society are fundamentally changing due to the emergence of Big Data. Some people see it as a natural and positive change, while others are more critical as they worry about the risks Big Data Analysis pose to Science and Society. We begin by surveying the debate concerning potential changes in the way scientific research is or will be conducted, and move on to the societal effects of Big Data Analysis.

3.1 Big Data Analysis and the Scientific Method

A few prominent researchers have recently suggested that there is a revolution underway in the way scientific research is conducted. This argument has three main points:

- Traditional statistics will not remain as relevant as it used to be,
- Correlations should replace models, and
- Precision of the results is not as essential as it was previously believed to be.

These arguments, however, are countered by a number of other scientists who believe that the way scientific research is conducted did not and should not change as radically as advocated by the first group of researchers. In this section, we look at the arguments for and against these statements.

Arguments in support of the Big Data revolution

The four main proponents of this vision are Cukier, Mayer-Shoenberger, Anderson and Pentland [3, 47, 52]. Here are the rationales they give for each issue:

Traditional Statistics Will Not Remain as Relevant as It Used to Be: With regard to this issue, Cukier and Mayer-Schoenberger [47] point out that humans have always tried to process data in order to understand the natural phenomena surrounding them and they argue that Big Data Analysis will now allow them to do so better. They believe that the reason why scientists developed Statistics in the 19th century was to deal with small data samples, since, at that time, they did not have the means to handle large collections of data. Today, they argue, the development of technology that increases computer power and memory size, together with the so-called "datafication" of society makes it unnecessary to restrict ourselves to small samples.

This view is shared, in some respect, by Alex Pentland who believes that more precise results will be obtainable, once the means to do so are derived. He bases his argument on the observation that Big Data gives us the opportunity not to aggregate (average) the behaviour of millions, but instead to take it into consideration at the micro-level [52]. This argument will be expanded further in a slightly different context in the next subsections.

Correlations Should Replace Models: This issue was advocated by Anderson in his article provocatively titled "The End of Theory" [3] in which he makes the statement that theory-based approaches are not necessary since "with enough data the numbers speak for themselves". Cukier and Mayer-Schoenberg agree as all three authors find that Big Data Analysis is changing something fundamental in the way we produce knowledge. Rather than building models that explain the observed data and show what causes the phenomena to occur, Big Data forces us to stop at understanding how data correlates with each other. In these authors' views, abandoning explanations as to why certain phenomena are related or even occur can be justified in many practical systems as long as these systems produce accurate predictions. In other words, they believe that "the end justifies the means" or, in this case, that "the end can ignore the means". Anderson even believes that finding correlations rather than inducing models in the traditional scientific way is more appropriate. This, he believes, leads to the recognition that we do not know how to induce correct models, and that we simply have to accept that correlations are the best we can do. He further suggests that we need to learn how to derive correlations as well as we can since, despite them not being models, they are very useful in practice.

Precision of the Results Is Not as Essential as It Was Previously Believed to Be: This issue is put forth by Cukier and Mayer-Schoenberger who assert that "looking at vastly more data (...) permits us to loosen up our desire for exactitude" [47]. It is, once again, quite different from traditional statistical data analysis, where samples

had to be clean and as errorless as possible in order to produce sufficiently accurate results. Although they recognize that techniques for handling massive amounts of unclean data remain to be designed, they also argue that less rigorous precision is acceptable as Big Data tasks often consists of predicting trends at the macro level. In the Billion Price Project, for example, the retail price index based on daily sales data in a large number of shops is computed from data collected from the internet [12]. Although these predictions are less precise than results of systematic surveys carried out by the US Bureau of Labour Statistics, they are available much faster, at a much lesser cost and they offer a sufficient accuracy for the majority of users.
The next part of this subsection considers the flip-side of these arguments.

Arguments in denial of the Big Data revolution

There have been a great number of arguments denying that a Big Data revolution is underway, or at least, warning that the three main points just discussed are filled with misconceptions and errors. The main proponents of these views are: Danah Boyd and Kate Crawford, Zeynep Tufekci, Tim Harford, Wolfgang Pietsch, Gary Marcus and Ernest Davis, Michael Jordan, David Ritter, and Alex Pentland (who participates in both sides of the argument). Once again, we examine each issue separately.
Traditional Statistics Will Not Remain as Relevant as It Used to Be: The point suggesting a decline in the future importance of traditional Statistics in the world of Big Data Analysis raises three sets of criticisms. The first one comes with a myriad of arguments that will now be addressed:

- Having access to massive data sets does not mean that there necessarily is a sufficient amount of appropriate data to draw relevant conclusions from without having recourse to traditional statistics tools
  In particular,
  - **Sample and selection biases will not be eliminated:** The well-known traps of traditional statistical analysis will not be eliminated by the advent of Big Data Analysis. This important argument is made by Danah Boyd and Kate Crawford as well as Tim Harford and Zeynep Tufekci. Tufekci, in particular, looks at this issue in the context of Social Media Analysis [64]. She notes, for example, that most Social Media research is done with data from Twitter. The reasons are that Twitter data is accessible to all (Facebook data, on the other hand, is proprietary) and has an easy structure. The problem with this observation is that not only is Twitter data not representative of the entire population, but by the features it presents it forces the users to behave in certain ways that would not necessarily happen on different platforms.
  - **Careful Variable Selection is still warranted:** The researchers that argue that more data is better and that better knowledge can be extracted from large data sets are not necessarily correct. For example, the insights that can be extracted from a qualitative study using only a handful of cases and focusing on a few carefully selected variables may not be inferable from a quantitative study using thousands of cases and throwing in hundreds of variables simultaneously, see, e.g. Tim Harford's essay [34].

  - **Unknowns in the data and errors are problematic:** These are other problems recognized by both Boyd and Crawford and Tufekci [13, 14, 64]. An example of unknowns in the data is illustrated as follows: a researcher may know who clicked on a link and when the click happened, based on the trace left in the data, but he or she does not know who saw the link and either *chose* not to click it or *was not able* to click it. In addition, Big Data sets, particularly those coming from the Internet, are messy, often unreliable, and prone to losses. Boyd, Crawford and Tufekci believe that these errors may be magnified when many data sets are amalgamated together. Boyd and Crawford thus postulate that the lessons learned from the long history of scientific investigation, which include asking critical questions about the collection of data and trying to identify its biases, cannot be forgotten. In their view, Big Data Analysis still requires an understanding of the properties and limits of the data sets. They also believe that it remains necessary to be aware of the origins of the data and the researcher's interpretation of it. A similar opinion is, in fact, presented in [44, 51].
  - **Sparse data remains problematic:** Another very important statistical limitation, pointed out by Marcus and Davis in [44], is that while Big Data analysis can be successful on very common occurrences it will break down if the data representing the event of interest is sparse. Indeed, it is not necessarily true that massive data sets improve the coverage of very rare events. On the contrary, the class imbalance may become even more pronounced if the representation of common events increases exponentially, while that of rare events remains the same or increases very slowly with the addition of new data.

- **The results of Big Data Analysis are often erroneous:** Michael Jordan pulled the alarm on Big Data Analysis by suggesting that a lot of results that have been and will continue to be obtained using Big Data Analysis techniques are probably invalid. He bases his argument on the well-known statistical phenomenon of spurious correlations. The more data is available, the more correlations can be found. With current evaluation techniques, these correlations may look insightful, when, in fact, many of them could be discarded as white noise [2]. This observation is related to older statistical lessons on dealing with other dangers, such as the multiple comparison problems and false discovery.
- **Computing power has limitations:** [24] points out that even if computational resources improve, as the size of the data sets increases, the processing tools may not scale up quickly enough and the computations necessary for data analysis may quickly become infeasible. This means that the size of the data sets cannot be unbounded since even if powerful systems are available they can quickly reach their limit. As a result, sampling and other traditional statistical tools are not close to disappearing.

Correlations Should Replace Models: This issue is, once again, countered by three arguments:

- **Causality cannot be forgone:** In their article, Boyd and Crawford completely disagree with the provocative statement by Chris Anderson that Big Data Analysis will supersede any other type of research and will lead to a new theory-free perspective. They argue, instead, that Big Data analysis is offering a new tool in the scientific arsenal and that it is important to reflect on what this new tool adds to the existing ones and in what way it is limited. In no way, do they believe, however, that Big Data analysis should replace other means of knowledge acquisition since they believe that causality should not be replaced by correlations. Each has their place in scientific investigation. A similar discussion concerning the need to appreciate causality is expressed by Wolfgang Pietsch in his philosophical essay on the new scientific methodology [51]
- **Correlations are not always sufficient to take action:** In his note entitled "When to act on a correlation, and when not to", Ritter considers the dilemma of whether one can intervene on the basis of discovered correlations [53]. He recommends caution while taking actions. However, he also claims that the choice of acting or not depends on balancing two factors: (1) confidence that the correlation will reoccur in the future and (2) trade-off between risk and reward of acting. Following this, if the risk of acting and being wrong is too high, acting on strong correlations may not be justified. In his opinion, confidence in a correlation is a function of not only the statistical frequency but also the understanding of what is causing that correlation. He calls it the "clarity of causality" and shows that the fewer possible explanations there are for a correlation, the higher the likelihood that the two events are really linked. He also says that causality can matter tremendously as it can drive up the confidence level of taking action. On the other hand, he also distinguishes situations where, if the value of acting is high, and the cost of wrong decisions is low, it makes sense to act based on weaker correlations. So, in his opinion a better understanding of the dynamics of the data and working with causality is still critical in certain conditions, and researchers should better identify situations where correlation is sufficient to act on and what to do when it is not.
- **Big Data Analysis will allow us to understand causality much better:** Unlike Anderson and Cukier and Mayer-Schoenberger, Alex Pentland does not believe in a future without causality. On the contrary, in line with his view that Big Data Analysis will lead to more accurate results, he believes that Big Data will allow us to understand causalities much more precisely than in the past, once new methods for doing so are created. His argument, as seen earlier, is that up to now, causalities were based on averages. Big Data, on the other hand, gives us the opportunity not to aggregate the behaviour of millions, but instead to take it into consideration at the micro-level [52].

Precision of the Results Is Not as Essential as It Was Previously Believed to Be: This argument in favour of decreasing the rigour of the results is countered by two arguments as follows: