INFORMATION SYSTEMS, WEB AND PERVASIVE COMPUTING SERIES

Data Analytics and Big Data

Soraya Sedkaoui

ISTE WILEY

Data Analytics and Big Data

To “Ben M’hidi”
My idol and the soul of my homeland

Data Analytics and Big Data

Soraya Sedkaoui

ISTE

WILEY

First published 2018 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

Library of Congress Control Number: 2018936255

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN 978-1-78630-326-4

Contents

Acknowledgments

"No guide, no realization".

It is true that writing a book needs time, patience and motivation in equal measures. However, the use of analytics, the application of algorithms and uncovering the hidden patterns behind the data available today have always excited me. When we consider the opportunities offered by the big data universe, the power of analytics and what may be revealed by each byte of data, the effort involved to write this book must be doubled.

I would be remiss if I did not mention the excellent advice and additional motivation that I received from Professor Hans-Werner Gottinger and Professor Jean-Louis Monino, who helped me to shape my ideology on how big data analytics can be applied to generate value. Their guidance and useful advice helped me to pursue my ultimate dream of writing a book. Thank you for everything!

I must also acknowledge my beloved family: my mother, as I would not be doing this if it was not for her and the drive to make her proud of me; my sisters and brother (Saliha, Nadia, Zahra and Kamel) and, with special attention, Manel and Zaki, for their continuous encouragement, support and help in every step that I take. They provide me with the strength that I need to go forward. I am very

grateful to have such a wonderful and supportive family; they are great people and without them, this book may not have been written.

Also, my sincere thanks to my friends who support me and understand that I do not have much time but I still count on the love and support that they have given me throughout my career and the development of this book.

Soraya

Preface

> “If you can look into the seeds of time,
> And say which grain will grow and which will not,
> Speak then to me, who neither beg nor fear
> Your favors nor your hate”.
>
> Shakespeare, Macbeth, Act I, Scene III, 59–62.

This book treats the roots and the fruits of the movement that marks, affects and transforms any part of business and society. It is about the large amounts of data (the seeds of our time) that we are sowing and creating by simple contact with our connected objects or simple use of advanced IT tools and the value generation that we have to derive and reap, as Shakespeare suggests, through sophisticated methods and advanced tools.

At the time of reading this book, you have to know that more different types of data will be produced. It is no longer about the word “big”, but it is more about how to handle this “big” amount of structured and unstructured data, which cannot be managed with traditional tools, and deal with its diversity and velocity to generate value.

Therefore, this book is about "big data analytics", which are probably nothing new in reality but have become one of the most exciting fields of our time. This exciting field opens the way to new opportunities that have significantly changed the business playground.

We have probably noticed that "big" companies such as Google, Facebook, Apple, Amazon, IBM, Netflix and many other companies invest continuously in big data and analytics applications in order to take advantage of every data byte. Many companies have realized that knowledge is power, and in order to get this power they have to gather its source, which is data, and make sense of it.

However, with great power comes great responsibility! Thus, the mission of this book is to provide the reader with the different concepts and applications behind big data analytics, those that are necessary and most important in order to be familiar with the ways in which data analytics process and algorithms work, and how to use them.

Every chapter of this book is meant for readers who are looking to discover the importance of analytics tools and the pertinence of algorithm applications, and who have a critical vision toward how knowledge or this "power" is derived from data.

So, if you want to become a data analysis practitioner or a better problem solver, or even if you are considering a career in big data and joining the analytics arena, then this book is for you! If you are familiar with big data analytics techniques and Machine Learning (ML) algorithm applications and you want to enrich your knowledge and gain more insights into how it works, then this book will help you to put your knowledge into practice.

Also, if you are a novice in this field and you are seeking to developing your analytics ability, then this book is for you, too! This book will provide you a complete overview related to this context. So, do not worry, because even if you are completely new to the big data universe, analytics techniques and ML algorithm applications, this book will change the way that you think about it. You will realize at the end of this book that it can be an exciting field for you, too.

By writing this book, I want to share my knowledge in the hope that the reader will embrace the opportunity offered by this practical exciting context and focus on its applications. The necessary theoretical concepts behind big data analytics and ML will be simplified in order for the reader to understand how make sense out of data.

Before we dive into this universe, I say: “may the big data analytics power and ML algorithms’ relevance be with you”!

Dr. Soraya SEDKAOUI

March 2018

Introduction

It is quite natural for academics who are continuously passionate to publish and share their knowledge, and to want to always create something from scratch that is their own fresh creation.

It is true that writing a book is a huge investment in time and energy, but the most essential thing is to do a great work. This book is an experiment in not starting from scratch, as it is instead a “redesigning” of my previous works, which are related to the data analytics field.

The genesis of the idea for this book began in early 2017, after I was lucky enough to be part of many teaching programs, research endeavors and conferences. In that time, I told myself that it was time to write the book focused on “big data analytics”.

While writing this book, I suggest that the reader must have some basic concepts and methods related to statistics, linear algebra and mathematics. But, you do not have to worry because even if you have forgotten most or some of it, this book will help you to refresh your understanding of these concepts and methods.

So, if you want to understand big data analytics, its complexity, promises and applications of its models and mechanisms, as well as machine learning algorithms, then I tell you, whoever are you (student, manager, academic, etc.), welcome to this book!

But, remember that "I can only show you the door. You're the one that has to walk through it". (Morpheus, *The Matrix*)

Why this book?

As a trend that has emerged around the business context, a first reflex is to think that data analytics is like a fast and furious phenomenon or even a kind of magic ball that can predict all kinds of things with extraordinary precision. In the case of Google, Facebook, Amazon, as well as banks and insurers, the constitution of huge databases gives an increasingly central place to "big data analytics".

Big data analytics has become an extremely important and challenging problem in disciplines such as computer science, biology, medicine, finance and homeland security. As massive amounts of data are available for analysis, scalable integration techniques become important.

Nowadays, companies are starting to realize the importance of using more data in order to support decision for their strategies. It was said and proved through case studies that "more data usually beats better algorithms".

Data sizes have been growing exponentially within many companies. Facing this size of data – meta-tagged piecemeal, produced in real time, and arriving in continuous streams from multiple sources – and analyzing the data, to spot patterns and extract useful information, is still harder.

This includes the ever-changing landscape of data and their associated characteristics, evolving data analysis paradigms, challenges of computational infrastructure, data sharing and data access, and – crucially – our ability to integrate datasets and their analysis toward an improved understanding.

New forms of methods and technologies are required to analyze and process these data. This need has motivated the development of big data analytics and machine learning algorithms in this book.

The objective is to familiarize anyone who is curious to have an overview of big data analytics as a tool for addressing and applying new analytics methods and algorithms of machine learning, in order to process data and make more intelligent decisions.

Whom is this book for?

This book provides a basic introduction to big data analytics, data science and machine learning algorithms, which are being adopted and used more frequently, especially in businesses that are looking for new methods to develop smarter capabilities and tackle challenges in the dynamic processes.

It will help those who are interested in developing a broad picture of the current context characterized by big data analytics and machine learning, and enable them to recognize the possible trajectories of future developments. It will provide for those seeking to build a common set of concepts, terms, references, methods, applications and approaches in this area.

Organization of the book

"Paths are made by walking".

Franz Kafka

The concepts behind big data analytics are actually nothing new. Organizations have always used descriptive, predictive and perspective analytics (business intelligence), and academics and researchers have been using data to analyze phenomena for many years. However, the amount of data available today and the emergence of the big data age in the early years of this decade, which impose many challenges, are changing the data analytics arena.

The challenge, therefore, lies in the ability to extract value from the volume of data produced in real-time continuous streams in multiple forms and from multiple sources. In other words, the key to exploring data and uncovering secrets from it, is to find and develop applicable ways in which to extract knowledge that can conduct decision-making processes and business strategies.

This is what this book will explore by highlighting the contents in three parts.

The first part discusses the general context of the big data area and presents the corresponding state of the art. It offers, in Chapters 1 and 2, the general theoretical background and framework necessary to understand the rest of this book. This first part will cover the key challenges and benefits of big data. It gives a platform to precede to different big data-related concepts and how this phenomenon is changing business opportunities.

The second part contains three chapters, (Chapters 3–5), dedicated to the data analytics process, which mainly focuses on how we can make sense of data, and the essential tools and technologies for organizing, analyzing and benefiting from big data. It illustrates the power of advanced analytics and its wide range of applications by showing how it can be applied in order to solve fundamental data analysis tasks.

The three chapters of the third part (Chapters 6–8) introduce the main subareas of artificial intelligence (AI) and machine learning (ML). They discuss the essential ML algorithm families that can be used to tackle various problem tasks by giving a machine the ability to learn from data in order to better guide the model building paths.

Glossary

In order to attain a basic understanding of what big data analytics entails, it is necessary to provide and review the terms that shape a framework related to this field. This section introduces the concepts that are most associated with "big data analytics".

– **Algorithm:** A set of computational rules to be followed to solve a mathematical problem. More recently, the term has been adopted to refer to a process to be followed, often by a computer.

– Amazon Web Services (AWS): This is a comprehensive, evolving cloud computing platform provided by Amazon.com. Web services are sometimes called cloud services or remote computing services. The first AWS offerings were launched in 2006 to provide online services for websites and client-side applications.

– Analytics: This has emerged as a catch-all term for a variety of different business intelligence (BI) and application-related initiatives. For some, it is the process of analyzing information from a particular domain, such as Website Analytics. For others, it is applying the breadth of BI capabilities to a specific content area (for example sales, service, supply chain and so on). In particular, BI vendors use the "analytics" moniker to differentiate their products from the competition. Increasingly, "analytics" is used to describe statistical and mathematical data analysis that clusters, segments, scores and predicts what scenarios are most likely to happen. Whatever the use cases, "analytics" has moved deeper into the business vernacular.

Analytics has garnered a burgeoning interest from business and IT professionals looking to exploit huge amounts of internally generated and externally available data.

– **Artificial intelligence:** The theory and development of computer systems able to perform tasks that traditionally have required human intelligence.

– **Big data:** A generic term that designates the massive volume of data that is generated by the increased use of digital tools and information systems. The term big data is used when the amount of data that an organization has to manage reaches a critical volume that requires new technological approaches in terms of storage, processing and usage. Volume, velocity and variety are usually the three criteria used to qualify a database as "big data".

– **Business intelligence** (BI): This is an umbrella term that includes the applications, infrastructure and tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance.

– **Cloud computing:** The National Institute of Standards and Technology (NIST) definition of cloud computing: "Cloud Computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications), and services that can be rapidly provisioned and released with minimal management effort or service provider interaction". This term designates a set of processes that use computational and/or storage capacities from remote servers connected through a network, usually the Internet. This model allows access to the network on demand. Resources are shared and computational power is configured according to requirements.

– **Cluster analysis:** A statistical technique whereby data or objects are classified into groups (clusters) that are similar to one another but different from data or objects in other clusters.

– **Computer science:** Computer science is the study of how to manipulate, manage, transform and encode information.

– **Customer relationship management (CRM):** This is a business strategy that optimizes revenue and profitability while promoting

customer satisfaction and loyalty. CRM technologies facilitate the implementation of a strategy, and make it possible to identify and manage customer relationships, in person or virtually. CRM software provides functionality to companies in four segments: sales, marketing, customer service and digital commerce.

– **Cyber security:** This is also known as computer security or IT security; it is involved in the protection of computer systems from the theft or damage of hardware, software or the information on them, as well as from disruption or misdirection of the services they provide.

– **Data:** This term comprises facts, observations and raw information. Data itself have little meaning if it is not processed.

– **Data analysis:** This is a class of statistical methods that makes it possible to process a very large volume of data and identify the most interesting aspects of its structure. Some methods help to extract relations between different sets of data, and thus draw statistical information that makes it possible to describe the most important information contained in the data in the most succinct manner possible. Other techniques make it possible to group data in order to identify its common denominators clearly, and thereby understand them better.

– **Data mining:** This practice consists of extracting information from data with the objective of drawing knowledge from large quantities of data through automatic or semiautomatic methods. Data mining uses algorithms drawn from disciplines as diverse as statistics, artificial intelligence and computer science in order to develop models from data, that is, in order to find interesting structures or recurrent themes according to criteria determined beforehand, and to extract the largest possible amount of knowledge useful to companies. It groups together all technologies capable of analyzing database information in order to find useful information and possible significant and useful relationships within the data.

– **Data science:** This is a new discipline that combines elements of mathematics, statistics, computer science and data visualization. The objective is to extract information from data sources. In this sense, data science is devoted to database exploration and analysis. This

discipline has recently received much attention due to the growing interest in big data.

– **Deep learning:** This is also known as deep structured learning or hierarchical learning; it is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms.

– **Exploratory data analysis (EDA):** In statistics, EDA is an approach of analyzing datasets to summarize their main characteristics, often with visual methods.

– **Garbage in, garbage out (GIGO):** In the field of computer science or information and communications technology, this refers to the fact that computers, since they operate through logical processes, will unquestioningly process unintended, even nonsensical, input data ("garbage in") and produce undesired, often nonsensical, output ("garbage out"). The principle applies to other fields as well.

– **Hadoop:** Big data software infrastructure that includes a storage system and a distributed processing tool.

– **Information:** This consists of interpreted data and has discernible meaning. It lies in descriptions and answers questions like "Who?" "What?", "When?" and "How many?"

– **Innovation:** Innovation can refer to something new or to a change made to an existing product, idea or field.

– **Internet of Things (IoT):** The internetworking of physical devices, vehicles, buildings and other items embedded with electronics, software, sensors, actuators and network connectivity that enable these objects to collect and exchange data and send, receive and execute commands. According to the Gartner group, IoT is the network of physical objects that contain embedded technology to communicate and sense or interact with their internal states or the external environment.

– **Knowledge:** This is a type of know-how that makes it possible to transform information into instructions. Knowledge can either be obtained through transmission from those who possess it or by extraction from experience.

– **Machine learning:** A method of designing a sequence of actions to solve a problem that automatically optimizes through experience and with limited or no human intervention.

– **Machine-to-machine (M2M):** Communications is used for automated data transmission and measurement between mechanical or electronic devices. The key components of an M2M system are field-deployed wireless devices with embedded sensors or RFID-wireless communication networks with complementary wireline access. This includes, but is not limited to cellular communication, Wi-Fi, ZigBee, WiMAX, wireless LAN (WLAN), generic DSL (xDSL) and fiber to the x (FTTx).

– **MapReduce:** This is a programming model or algorithm for the processing of data using a parallel programming implementation and was originally used for academic purposes associated with parallel programming techniques.

– **Natural language processing (NLP):** An interdisciplinary field of computer science, artificial intelligence and computation linguistics that focuses on programming computers and algorithms to parse, process and understand human language.

– **Nowcasting:** Nowcasting is the prediction of the present, the very near future and the very recent past in economics. The term is a contraction for *now* and *forecasting* and has been used for a long-time in meteorology. It has recently become popular in economics as standard measures used to assess the state of an economy. Nowcasting models have been applied in many institutions, in particular Central Banks, and the technique is used routinely to monitor the state of the economy in real time.

– **Open data:** This term refers to the principle according to which public data (that are gathered, maintained and used by government bodies) should be made available to be accessed and reused by citizens and companies.

– **Open innovation:** This is defined as the increased use of information and knowledge sources external to the company, as well as the multiplication of marketing channels for intangible assets with the purpose of accelerating innovation.

– **Open source:** A designation for a computer program in which underlying source code is freely available for redistribution and modification.

– **Scalability:** The measure of a system's ability to increase or decrease in performance and cost in response to changes in application and system processing demands. Enterprises that are growing rapidly should pay special attention to scalability when evaluating hardware and software.

– **Smart data:** The flood of data encountered by ordinary users and economic actors will bring about changes in behavior, as well as the development of new services and value creation. These data must be processed and developed in order to become "smart data". Smart data is the result of analysis and interpretation of raw data, which makes it possible to effectively draw value from it. It is, therefore, important to know how to work with the existing data in order to create value.

– **Statistical inference:** This is defined as the process of deducing properties of an underlying distribution through the analysis of data. Inferential statistical analysis infers properties about a population: this includes testing hypotheses and deriving estimates. The population is assumed to be larger than the observed dataset; in other words, the observed data are assumed to be sampled from a larger population.

– **Supervised learning:** A supervised learning algorithm applies a known set of input data and drives a model to produce reasonable predictions for responses to new data. Supervised learning develops predictive models using classification and regression techniques.

– **Terabyte:** A unit of data storage, equal to one trillion (10^{12}) bytes, or 1,000 gigabytes.

– **Text mining:** Equivalent to text analytics, text mining is the process of deriving information from text. Text mining usually involves the process of structuring the input text; deriving patterns within the structured data, and finally evaluation and interpretation of the output.

– **Unsupervised learning:** Unsupervised learning identifies hidden patterns or intrinsic structures in the data. It is used to draw

conclusions from datasets composed of labeled unacknowledged input data.

– **Web 2.0:** This term designates the set of techniques, functions and uses of the World Wide Web that have followed the original format of the web. It concerns, in particular, interfaces that allow users with little technical training to appropriate new web functions. Internet users can contribute to information exchanges and interact (share, exchange, etc.) in a simple manner.

– **World Wide Web:** This is a hypertext-based global information system that was originally developed at the European Laboratory for Particle Physics in Geneva. It is a subset of the Internet, technically defined as the community on the Internet where all documents and resources are formatted using Hypertext Markup Language (HTML). HTML, and the related Hypertext Transport Protocol (HTTP); making it easy to find and view data and documents stored on computers connected to the Internet. HTML creates the links ("hyperlinks") that enable the user to move among many web documents with the click of a mouse.

– **XML:** XML Extensible Markup Language is a markup language that defines a set of rules for encoding documents in a format that is both human readable and machine readable.

– **Yottabytes:** A unit of data storage that is equal to 1 sextillion (102^4) bytes.

– **Zettabyte:** A unit of data storage that is equal to 1 sextillion (10^{21}) bytes, 1 trillion gigabytes or 1 billion terabytes.

PART 1

Towards an Understanding of Big Data: Are You Ready?

1

From Data to Big Data: You Must Walk Before You Can Run

> "If I had to express my views about the digital future – that of Europe or indeed, of the whole world – I could do it with one word: data".
>
> Andrus Ansip, Vice-President, Digital Single Market, (2015)

1.1. Introduction

Big data is the revolutionary word in today's world because of its influence on several domains. Somehow, the debates that have been emerging over the last few years around big data are very similar to those that took place about the "Web" in the early 1990s. After a long and active discussion phase in the literature, big data entered a phase of use by many companies.

One of the reasons big data has become so popular is that the availability of data has improved. This phenomenon has radically changed the way data are collected and analyzed since it introduces new issues concerning the *volume (how much?)*, *velocity (at what speed?)* and *variety (how diverse?)* of data available today.

Data Analytics and Big Data, First Edition. Soraya Sedkaoui.

In this chapter, we will cover and discuss the basic concepts that lie behind the big data age in order to highlight the importance of working with data. Deeper concepts will be introduced later in the following chapters when we need them. For the moment, the goal of this chapter is to give an overview of the big data domain without yet diving into big data analytics processes.

1.2. No analytics without data

Nowadays, it does not take much to convince managers or decision-makers alike of the importance of data for their business activities because most business activities are associated with the use, understanding and exploitation of data.

For example, 1, 10, 28 and 80 are just numbers, but if someone says: "10 thousand YouTube videos have generated more than 1 billion views", "men spend 28% more money online than women" and "almost 80% of time spent on social media platforms happens on mobile". Then, 1, 10, 28 and 80 become a piece of information that can help us to know more about certain events.

Data are a collection of facts, such as numbers, words, measurements, observations or even just descriptions of things, which give more information about an individual, an object or an observation.

According to the Oxford dictionary, data are defined as:

> "the quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media".

Understanding individual/object/observation requires us to explore the term "data".

Data are, therefore, a form of wealth, and exploiting it results in an essential competitive advantage for an ever-tougher context. In the big

data age, we will come across many different types of data, and each of them requires different tools and techniques.

Now that you have, at least, an idea about what data is, you may be wondering where these data come from? We will quickly give an overview of the usual type of data encountered.

1.2.1. *Databases*

Databases are the main source of data. There are a plethora of technologies (from Hadoop to SQL) to ensure the recovery and the storage of data. These databases may include different types of information (the logs of a web server, banking transactions, etc.).

	A	B	C	D	E
1	**ID**	**Gender**	**Country**	**Age**	**Degree**
2	1567752	Female	France	25	Master
3	1567624	Male	Germany	28	Ph.D
4	1567301	Male	USA	30	Master
5	1567112	Male	Italy	33	4-year college
6	1567267	Female	UK	28	Professional
7	1567058	Male	USA	24	4-year college
8	1567532	Female	Russia	31	4-year college
9	1567450	Male	Spain	27	Master
10	1567743	Female	Italy	24	4-year college
11	1567713	Female	UK	29	Ph.D
12	1567981	Male	Netherlands	27	Professional
13	1567730	Male	France	34	4-year college
14	1567911	Female	Spain	29	4-year college

Figure 1.1. *Example of structured data (in Excel table)*

1.2.2. *Raw data*

Other raw data, often more complex and requiring specific preprocessing to make them manipulable by the algorithms, can serve as sources for a modeling problem.

1.2.3. *Text*

All text written in natural language by humans is thus another main source of data. This includes all types of text that can be thought of naturally (books, articles, etc.), but also other types of text such as HTML code or DNA sequences.

Figure 1.2. *Example of text analytics*

1.2.4. *Images, audios and videos*

Audio, images and videos are data types that pose specific challenges to a data scientist. Images, audio and videos are also helpful to capture as a source of data. Many images, audios and videos (classified by type or other) are available to process. Image, audio and video processing is a field of research in its own right. For example, DeepMind succeeded at creating an algorithm that is capable of learning how to play video games.

1.2.5. *The Internet of Things*

Connected objects are another source of raw data, which retrieves a large amount of data through their sensors. The Internet of Things (IoT) contributes to double the size of the digital universe every 2 years, which could be 44,000 billion gigabytes in 2020, 10 times more than in 2013 [EMC 14].

The connected objects thus extend the scope of the Internet allowing any object, machine or living element to transmit information about its environment, and eventually be activated remotely.

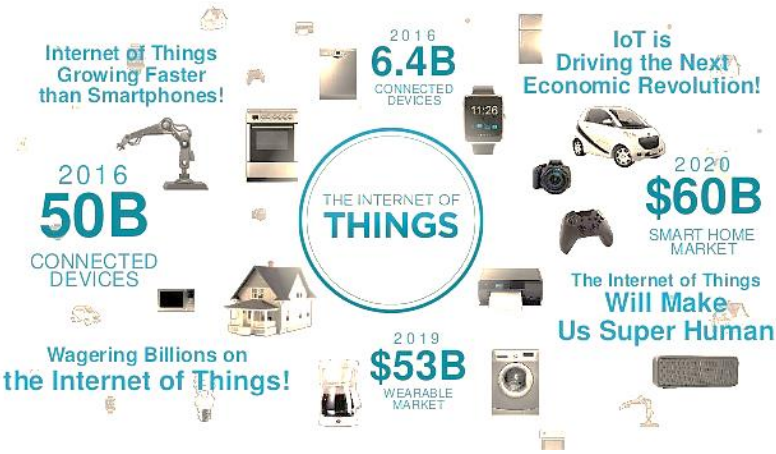


Figure 1.3. *Data generated by the IoT (source: https://steemit.com/technology/@adarshagni/the-internet-of-things-the-internet-world)*

1.3. From bytes to yottabytes: the data revolution

All revolutions have in common a complete change in the technical system that influences society. The typographic printing press developed by Gutenberg in 1450 was a symbol of the first industrial revolution. This first industrial revolution brought mechanization, centralized factories and industrial capitalists.

According to the historian Elizabeth Eisenstein, between 1457 and 1500 nearly 8 million books were printed, the equivalent of all that had been produced by the scribes of Europe since the foundation of Constantinople. As the stock of information in the world doubled in almost 50 years; it has doubled again in just 3 years.

Already, in 1941, there was talk of *Information Explosion* in the *Oxford English Dictionary* to evoke the exponential development of

data. But it is in the digital age that the term makes sense: according to the UN, more data was created in 2011 than in the whole history of humanity.

In 2011 alone, 1.8 zettabytes of data were created [GAN 11]. According to IBM, the proliferation of web pages, image and video applications, social networks, mobile devices, apps, sensors and so on is able to generate more than 2.5 quintillion bytes per day to the extent that 90% of the world's data have been created over the few past years [CUK 13a, CUK 13b, DIE 14, FOS 17].

As large companies frequently disclose that their datasets are growing vaster ever more rapidly, the amount of digital information is predicted to increase 10-fold every 5 years [THE 10]. Many of these companies' datasets are within the petabytes range but soon they could reach exabytes or even zettabytes. Data sizes have been growing exponentially within many organizations.

In 1999, Wal-Mart, one of America's most important retail chains, had a database of 1,000 TB (i.e. 1,000,000 GB of data). In 2004, Wal-Mart claimed to have the largest data warehouse with 500 TB storages. By 2012, this quantity had grown to over 2.5 PB (2.5 million GB) of data. In 2008, eBay storage amounted to 8 PB. Three years later, the Yahoo warehouse totaled 170 PB [SED 16].

In 2010, Eric Schmidt, the CEO of Google at the time, estimated the size of the World Wide Web at roughly 5 million TB of data, while the largest storage cluster within a corporation such as Facebook had more than 100 PB of data in 2013. In addition, Facebook reported storing 300 PB of data at a rate of 600 TB per day in 2014, facing a threefold increase in 1 year only [VAG 14].

Gartner [GAR 17] estimates that connected objects will reach 20.4 billion by 2020. Objects that lead to a multiplication of data quantity result in a diversification of its uses. These data will gradually increase as the IoT develops.

This allows us to think about the legend of the wise "Sissa" in India. When King "Belkib" asked about the reward he desired, after

his invention, he asked to receive a grain of rice for the first square, two grains for the second, four grains for the third and so on. The king agreed, but he did not know that on the last square of the board he should drop 2^{63} grains, or more than 700 billion tons.

In their book *Race against the Machine*, Brynjolfsson and McAffee [BRY 11] referenced the fable of the chess and rice grains to make the point that "exponential increases initially look a lot like linear, but they are not. As time goes by – as the world move into the second half of the chessboard – exponential growth confounds our intuition and expectation".

Thus, currently, not only is the quantity of digitally stored data much larger, but the type of data is also very varied, because of the various new technologies [SED 16].

Individuals are putting more and more publicly available data on the web. Many companies collect information on their clients and their respective behavior. As such, many industrial and commercial processes are being controlled by computers. The results of medical tests are also being retained for analysis.

Units	Symbol	Value
1 byte	–	8 bites
1 kilobyte	KB	10^3 bytes
1 megabyte	MB	10^6 bytes
1 gigabyte	GB	10^9 bytes
1 terabyte	TB	10^{12} bytes
1 petabyte	PB	10^{15} bytes
1 exabyte	EB	10^{18} bytes
1 zettabyte	ZB	10^{21} bytes
1 yottabyte	YB	102^4 bytes

Table 1.1. *Data: from byte to yottabyte*

Financial institutions, companies, health service providers and administrations generate large quantities of data through their interactions with suppliers, patients, customers and employees. Beyond those interactions, large volumes of data are created through Internet searches, social networks, GPS systems, and stock market transactions.

However, companies that produce, manage and analyze vast sets of data on a daily basis now commonly use terms such as terabyte, petabyte, exabytes, zettabyte [SED 16] and even yottabytes.

1.4. Big data: definition

Data exist over time, it is not new, but what makes them so important is the rapid rate and different forms in which they have been produced in recent times, or what brings us to turn: *From data to Big data*.

Big data is created digitally and collected automatically. Berman [BER 13] identified six different mechanisms through which data can be founded:

1) the data are already collected in the course of normal activities and are waiting to be used. The data owner does not want to discover or to do anything new, but to do what it has always been doing better;

2) the data are already collected but new activities are supported by the data;

3) a business model is planned based on a big data resource. An example of this mechanism is data-intensive services such as Amazon;

4) a group of entities that have large data resources federate their data resources, for example hospital databases;

5) large amounts of data are collected and organized to benefit an organization and their user-clients. These projects require skills and vision;

6) big data resources are built from scratch. No data and no big data technologies existed before big data.

The growth of available data in terms of quantity, diversity and access speed has been enormous, giving way to the "3Vs" as part of many big data definitions.

Typically, volume, velocity and variety are used to characterize the key properties of big data:

– *volume* refers to the size of the data;

– *velocity* refers to the data provisioning rate and the time within which it is necessary to act on them;

– *variety* refers to the heterogeneity of data acquisition, data representation and semantic interpretation.

These three dimensions will be developed later, but now let us explore how big data is defined.

The popular definition of big data is that given by Gartner [GAR 13], which defines big data as:

> "An information asset whose volume is large, velocity is high, and formats are various".

Research firm McKinsey [MCK 11] also offers their interpretation of what big data is:

> "Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data – i.e., we don't define big data in terms of being larger than a certain number of terabytes (thousands of gigabytes). We assume that, as technology advances over time, the size of datasets that qualify as big data will also increase. Also note that the definition can vary by sector,

depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry. With those caveats, big data in many sectors today will range from a few dozen terabytes to multiple petabytes (thousands of terabytes)''.

1.5. The 3Vs model

The "3Vs", which are generally used to describe this phenomenon, explain how big data represents the arrival of technologies that allow for a whole new approach to data. Modern information technology, incremental computing power and online digitalization have opened up new options for utilizing automatically collected and stored data from various sources in multiple formats.

In this case, the traditional way of formatting information from transactional systems to make them available for "statistical processing" does not work in a situation where data is arriving in huge volumes from diverse sources, and where even the formats could be changing [SED 17].

Traditional analysis methods have been based largely on the assumption that we can work with data within the confines of their own computing environment. But the growth of the amounts of data is changing this paradigm, especially the progress in computational data analysis.

Big data has put a great challenge on the current statistical methodology and computational tools. With growing size typically comes a growing complexity of data structures, data pattern and the models needed to account for these patterns.

Working with big data and applying a series of data analysis techniques in today's world, which is becoming increasingly obsessed with big data more than virtually anything else [THE 17], requires strong multidisciplinary skills and knowledge of statistics,

econometrics, computer science, data mining, law and business ethics, etc.

In addition, understanding big data means dealing with data volumes that are significantly larger than those previously analyzed, at an incomparable speed, all while integrating a widely rich data variety. So, each of these "Vs" deserves some clarification.

Big data implies an enormous *volume* of data. This volume is regarded as the first dimension for set points from big data. However, this first V is the least operational and most variable depending on the sector and the organization. Today, we are talking about storing and analyzing exabytes (10^{18}) or even zettabytes (10^{21}), whereas just 10 years ago we were talking about megabytes (10^{6}) stored on floppy disks.

The volume of data is enormous and a very large contributor to the ever expanding digital universe is the IoT with sensors all over the world in all devices creating data every second.

This flow is massive and continuous besides the speed at which data are created currently. Every minute we upload a huge amount of video data on YouTube. In addition, billions of e-mails are sent, billions of photos are viewed and uploaded, thousands of tweets are sent and billions of queries on Google are performed.

This speed refers to the *velocity* at which the data are created, analyzed and stored. It means that the immediacy and instantaneousness of receiving or transmitting data by all of us and for several activities, is very important and compels companies to improve their reaction and anticipation velocity. The diversity of sources and formats of data represents a real technological challenge.

Akerkar [AKE 14] and Zicari [ZIC 14] argued that, based on the data life cycle (see Figure 1.4), the broad challenges of big data can be grouped into three main categories: data, process and management.

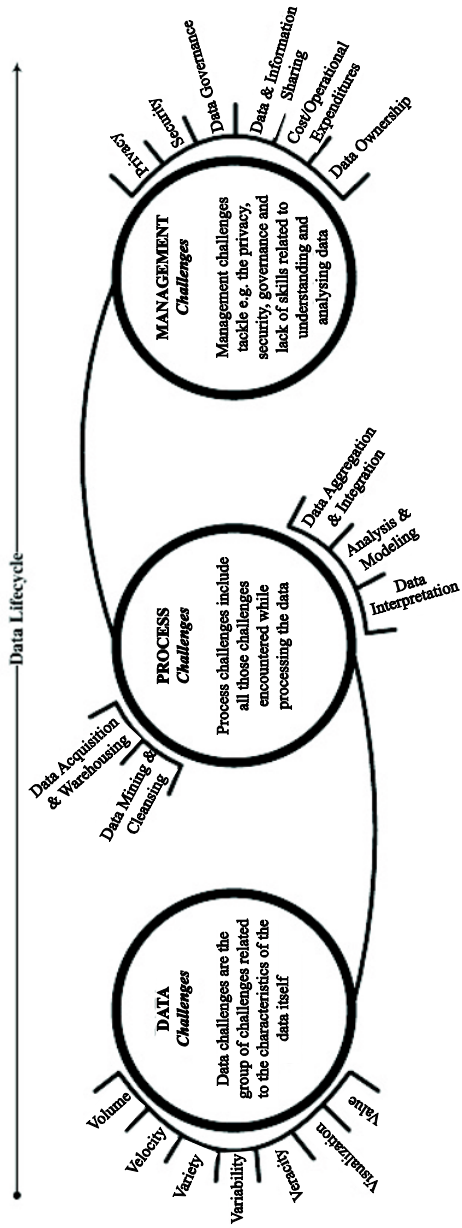


Figure 1.4. *Data life cycle*

Now data comes in the form of e-mails, photos, videos, monitoring devices, PDFs, audio, etc. There are many different types of data and each type requires different and specific types of analyses. Whatever is the format (text, images, photos, videos, etc.), a *variety* of data types need storage, mining and analyzing.

1.6. Why now and what does it bring?

The added value of big data is the ability to identify useful data and turn it into usable information by identifying patterns, exploiting new algorithms, tools and new project solutions. The idea behind the term "big data" is that it justifies that we are talking about revolution and not a simple development of data analysis.

What big data brings is the ability to process and analyze all types of data, in their original form, by integrating new methods and new ways of working.

It is the fact that these 3Vs change completely the way in which data are addressed, by putting it at the center of this transformation. Many companies' experiences, such as Netflix, Google, Uber, Facebook, Twitter, Nike and others, illustrate that data can deliver value in almost any area of business.

Turning data into information and then turning that information into knowledge (which is illustrated in Figure 1.5) remains a key factor for business success.

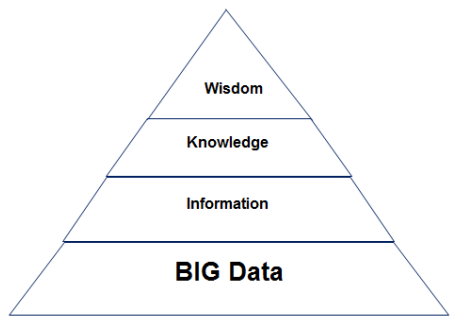


Figure 1.5. *Knowledge pyramid*

Overall, the approach seems to be simple: (1) we need data, (2) we need to know what we want to do with it and (3) how to do it. But, it is not. Why? Good question indeed! Here are some examples to better clarify.

– *Example 1*

Let us take a look at the image below:

(Source: Images via Google Image Search: https://fr.depositphotos.com/102864128/stock-illustration-flat-design-people-characters.html). For a color version of this figure, see www.iste.co.uk/sedkaoui/data.zip

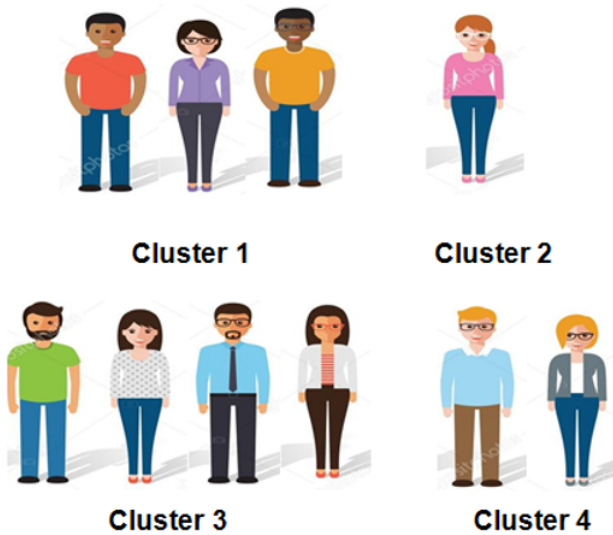
It is a collection of 10 individuals having different sex, age, job, nationality, etc. If we take a moment to categorize them by similarity into a number of groups, for example, start with grouping them by hair color. It is most likely that we split them into the following four groups (clusters):

– people with black hair in one cluster;

– red hair in another cluster;

– brown hair into one cluster;

– blond into one more cluster.



For a color version of this figure, see www.iste.co.uk/sedkaoui/data.zip

That was not too difficult, was it? It was amazing! Now, we could probably do the same with twice that number. And with a bit more time we could even categorize a hundred people.

Could you even categorize them by nationality or religion? Of course, yes! But how long would it take? And how many groups to seek? Besides, after obtaining the clusters, we also need to evaluate the validity and quality of the clusters.

But for a machine, however, it is an easy task. There are thousands of different possible ways to group these individuals into many meaningful clusters.

– *Example 2*

In the same way, if we look at all readers who may be interested in this book, we can group them according to many criteria, such as age

(from 18 to 70, for example), occupation (students, decision-makers, managers, researchers, etc.) and so on.

Concerning their age, it is defined by numbers, and it can vary from one reader to another. So, the data measuring this variable can be defined as follows:

$\text{Age}_{(1)} = 18, \text{Age}_{(2)} = 19, \text{Age}_{(3)} = 20, \ldots\ldots, \text{Age}_{(n-1)} = 69, \text{Age}_{(n)} = 70$

But, this is different from the number of readers R_x, when 'x' refers to the total number of readers and varies between 1 and m.

– R_1 is the reader 1;

– R_2 is the reader 2, etc.

If we say, for example, R_{10}, we are talking about the tenth reader. But, number 10 does not mean that he is 10.

By adding the other variables (occupation, degree, region, etc.) and combining them with each other, imagine how much information we can get.

We can even imagine the amount of information that can be generated by mixing these data, already known, with other data, for example, online data such as the number of like on Facebook, Twitter, etc.; the book's amount of web page visits; the number of appearances of the book on Google search; etc.

If we can imagine the increasing number of data (structured or not) for this book, then we can at least get an idea about the amount data and variety of companie's.

So, we can imagine how the rise of big data, and the series of analytical methods that can be applied, have significantly changed the business playground.

Congratulations! We are actually in the process of discovering that big data analytics is a challenging and exciting field that will certainly serve us.

These two examples can help us to realize that big data now occupies an important place in our life and mastering the fundamentals will make us a more critical "data scientist" thinker.

1.7. Conclusions

Recognizing the different types of data, their significance and where to look for them, understanding the importance of big data and realizing why a lot of attention has been paid to the "data revolution" were the objectives of this chapter.

But that is not all, because the emergence of the big data age is related not only to the several opportunities to investigate areas that were previously hard to examine, but also to its challenges and the way this phenomenon is changing businesses opportunities. So, follow along with Chapter 2 to further enrich our understanding about big data; its context, its challenges, its opportunities and the promises that it holds.

2

Big Data: A Revolution that Changes the Game

> "Upon this gifted age, in its dark hour, Rains from the sky a meteoric shower. Of facts . . . they lie unquestioned, uncombined. Wisdom enough to leech us of our ill. Is daily spun; but there exists no loom. To weave it into fabric".
>
> Edna St. Vincent Millay, 1939

2.1. Introduction

With the advent of digital technology and smart devices, a large amount of varied data is being generated every day. The volume of data will continue to grow and in a very real way. This widespread production of data has resulted in the age of big data that was discussed in Chapter 1.

The data we produce, as well as other data we accumulate, constitute a constant source of knowledge. Big data is therefore about collecting, analyzing and using data efficiently and quickly with new tools in order to gain a competitive advantage by turning data into knowledge and thus generate value.

Beyond what is big data and how it may impact the business context, this chapter highlights the other sides that have not been

Data Analytics and Big Data, First Edition. Soraya Sedkaoui.

addressed in Chapter 1. Through this chapter, we recall the context of big data, its importance in conducting decision-making, its challenges and the complementary role it plays in the creation of new opportunities for businesses.

2.2. Beyond the 3Vs

From the mid-1990s, at least, big data has increasingly appeared in publications and articles as a collective term for large amounts of data. The "3Vs" model" – this also forms the basis for the most used definition of big data (see Figure 2.1) – which is often cited today, dates back to 2001 and is based on a definition by Doug Laney, then an analyst with the market research firm Meta Group, which has since been taken over by its competitor Gartner. Ten years later, Gartner [GAR 12] gave this definition:

> "Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision-making, and process automation".

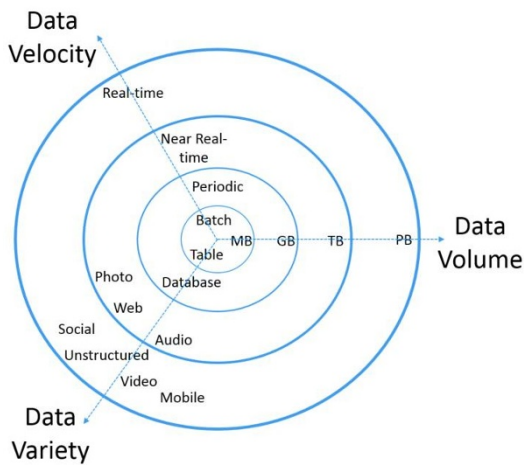


Figure 2.1. *The 3Vs that characterize big data (source: https://www.linguamatics.com/blog/big-data-real-world-data-where-does-text-analytics-fit)*

Big data goes beyond the three Vs mentioned. Some researchers integrate seven additional dimensions, and you know what? Each of these additional dimensions starts with V.

Table 2.1 shows the seven additional dimensions of big data with their characteristics.

	Vs	Characteristics
The seven additional Vs	Variability	Multitude of data dimensions resulting from multiple disparate data types and sources
	Veracity	Confidence or trust in the data, i.e. the provenance or reliability of the data source, its context and how meaningful it is for the analysis
	Validity	How accurate and correct are the data for their intended use?
	Vulnerability	The fact that a growing number of people are becoming switched on to the fact that their personal data are being gobbled up by the gigabyte, used to pry into their behavior and, ultimately, sell them things [MAR 16]
	Volatility	Describes how long data obtained in the original source is available and how long it should be stored
	Visualization	The way in which the results of data processing (information) are presented in order to ensure superior clarity
	Value	The end game after addressing the other Vs. It refers to the benefit of big data that can be gained through appropriate analysis

Table 2.1. *The seven additional Vs of big data*

These interrelated Vs, in addition to the 3Vs, are related to the quality of data and measure if the data have usefulness and are value relevant for their intended purpose. This is necessary because this can support decision making in an effective and efficient way, and data only matter when they are useful [FRI 16].

2.3. From understanding data to knowledge

The advent of IT tools has led to an explosion of data and has driven the computerization of production, service delivery and even the private sphere [WAL 16]. Data available today are very different from data that existed before. Data are collected from various sources. It can be in different infrastructures, such as cloud, or in different databases, such as rows, columns or files [MOO 15].

To the structured data, managed in traditional IT applications (ERP, CRM, etc.), have been added many other data, often called "unstructured data" or "semistructured data". These data mainly come from:

– *the web:* newspapers, social networks, e-commerce, indexing, documents, photos, video storage, linked data, etc.;

– *the Internet and connected objects:* sensor networks and Smart grid (cars, oil pipes, windmill turbines, etc.), call logs;

– *text data*: e-mail, news, Facebook feeds, documents, etc.;

– *location:* GPS and mobile phone as well as Wi-Fi connections makes time and location information a growing source of data;

– *science:* genomics, astronomy, subatomic physics;

– *commercial data:* transaction histories in a supermarket chain;

– *personal data*: medical files;

– *social network data:* social network sites such as Facebook, LinkedIn, Instagram, etc.;

– *public (open) data*.

Data	Data model	Examples
Structured	Relational data system	Databases
Semistructured	XML, CVS, logs	Web, logs, etc.
Unstructured	Text, image, video	E-mails, etc.

Table 2.2. *Structured, semistructured and unstructured data*

Previously, data were usually defined in a quantitative form (data being values that describe a measurable quantity) and qualitative form (describing of qualities or characteristics). These types of data will be considered as structured data because they require a simple transformation before issuing their meaning.

But, this is absolutely not the same case when we are faced with unstructured or semistructured data (web page, pdf, video, geolocation or sensors, etc.). Data can be divided into internal and external data as well as structured, semistructured and unstructured data.

We may be wondering what this has to do with the types of data that were previously presented. In fact, by analyzing such data, we apply a treatment aimed to reduce them to a sequence of numbers that can be interpreted by a "machine learning algorithm". Thus, it is easy to understand the need for powerful algorithms to process these types of complex data.

In addition to these categories, it is also helpful to look at data variety from a company's perspective: internal and external data. Along with capturing data from internal sales information and sensors, companies can also track public responses on Facebook, Twitter, or other social media.

When analyzed optimally, each type of data brings valuable insights that business leaders can use to make accurate and timely decisions. Business leaders also admit that "the role of digital

technologies is rapidly shifting, from being a driver of marginal efficiency to an enabler of fundamental innovation and disruption" [WOR 16].

These new types of data may be intended to enrich the types existing before; information is derived and then knowledge is produced, which is called the target paradigm of "knowledge discovery", described as a "knowledge pyramid" (Figure 1.5) where data lay at the base (see [ACK 89]).

Ackoff [ACK 96] defines data as symbols, information as data that are processed to be useful and knowledge as the application of data and information in order to have the ability to understand *why* and *how*.

According to Taylor [TAY 80], the value of information begins with data, which takes on value throughout its evolution until it achieves its objective and specifies an action to take during a decision. Information is a message with a higher level of meaning. It is raw data that a subject in turns transforms into knowledge through a cognitive or intellectual operation.

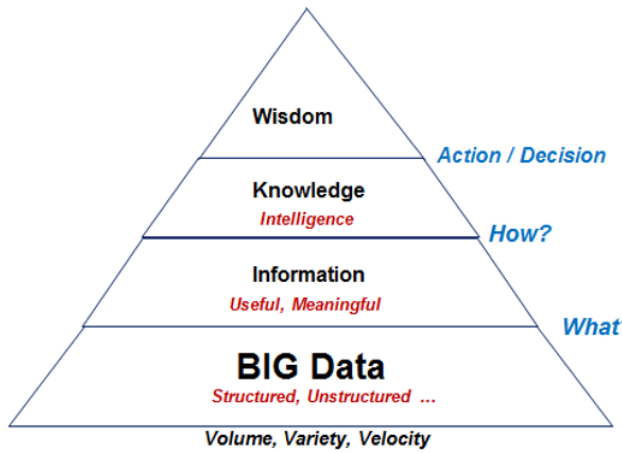


Figure 2.2. *Valuating data to extract knowledge*

Over the last decade, data has become the raw material of knowledge. Knowledge has a wider and deeper meaning than data or information. It is created from the use, analysis and productive utilization of data. This idea can be seen in the knowledge pyramid which illustrates that data are essential to build knowledge in order to make a good decision and generate value.

In fact, the emergence of big data has the potential to influence a company. According to Frizzo-Barker *et al*. [FRI 16], it has potential to change the thinking of companies about data infrastructure, business intelligence and analytics and information strategy [MCA 12].

By using big data, companies are able to measure significantly more about the business in their context and it is possible to make translations of that knowledge, which can improve the decision-making and the performance of the company [MCA 12].

2.4. Improving decision-making

Traditionally, the decision-making process is shaped on the model of limited rationality by Simon [SIM 77]: *Intelligence*, *modeling*, *choice* and *control*. With the exploitation of big data, this process has been made more complicated and has to improve. However, decision-making, business strategy development and the anticipation of change have always been dependent on the quantity and quality of data available.

The explosion of a phenomenal amount of data and the need to analyze puts forward the hierarchical model "from data to knowledge". The hierarchy is by nature a multidisciplinary endeavor [PIE 15]: computer scientists construct algorithms to manipulate and organize the data, aided by statisticians and mathematicians who instruct on the development and application of quantitative methodology.

Thus, database experts collect and warehouse the data, software designers write programs that apply the analytics algorithms to the data, engineers build electronics and hardware to implement the programming and subject-matter/domain experts – that is biologists, chemists, economists and social scientists – interpret the findings.

The most important asset of large volumes of data has to do with the fact that they make it possible to apply knowledge and create considerable value. However, in traditional models, key value creation activities can be described using the value chain [POR 85].

The value chain concept, which is primarily geared to the physical world, treats data as a supporting element rather than a source of value itself [RAY 95]. But, a correct utilization of these data in the decision-making process is not easy. The main challenge of using data is not in their collection, but in the choice of which data should be sought and how to make sense of them.

So, the process of decision-making begins when the top manager has to choose which data to look for, even before starting to collect data. Organizations need to use a structured view of data to improve their decision-making process. To achieve this structured view, they have to collect and store data, perform an analysis and transform the results into useful and valuable information.

Big data has the potential to aid in identifying opportunities related to decision-making in the intelligence phase of Simon's model, where the term "intelligence" refers to knowledge discovery. The intelligence phase is all about finding the occasions in which a decision should be made [SIM 97].

The major role of the intelligence stage is to identify the problem and collect relevant information [TUR 11], which would later be used in the next stages of the decision-making process. Also in this phase, the tools of "Business Intelligence" may be used to support the organization's discovery of opportunities for decision-making by providing advanced analytics and assuring data integration [POP 12].

Predictive modeling and analytics (which will be better developed in Part 2 of this book) can be of crucial importance for organizations if correctly aligned with their business process and needs, and can also lead to significant improvement of their performance and the quality of the decisions they make, thus increasing their business value (such as Amazon, eBay, Google, Facebook, etc.) [DAV 13, SIE 13].

Companies must first understand the potential value creation of connected devices and big data in their markets. Porter [POR 15] identifies four key capabilities of connected objects combined with big data:

1) *Monitoring*: Sensors placed on the objects provide information about their environment, the conditions of use and the objects' operation. The use of these data can be the source of new services, such as preventive medicine. These data can also be used indirectly to better consider the design of future objects, better segment the market and prices, or provide more effective customer service.

2) *Control*: Smart, connected products can be controlled through software embedded within them or that resides in the cloud. Users have an unprecedented ability to tailor product function and personalize interactions. Remote control of products increases employee safety and can reduce the number of employees needed.

3) *Optimization*: Algorithms and analytics can optimize product operation, capacity utilization and predictive maintenance.

4) *Autonomy*: Access to monitoring data, remote control, and optimization algorithms allow for product autonomy. This enables autonomous operation, self-coordination and self-diagnosis.

Indeed, when computers were first introduced in companies, it was in an attempt to tame the data.

Moreover, the type of analysis that is needed to be done on the data highly depends on the results to be obtained through decision-making. This can be done to:

1) incorporate massive data volumes in analysis;

2) determine upfront that big data is relevant.

So, we have two technical entities that have come together. First, there is big data for massive amounts of data. Second, there is advanced analytics, which is actually a collection of different tool types, including those based on predictive analytics, data mining,

statistics, clustering, data visualization, text analytics, artificial intelligence and so on [SHR 13, SIE 16].

The general aim of decision-making in the era of big data is to reduce problems to a scale that can be comprehended. Big data brings along with it some huge analytical challenges. Van Barneveld *et al*. [VAN 12] collate a number of definitions for the term analytics, concluding that analytics is understood as "data-driven decision-making".

The company embarking on the path of data-driven decision-making sees all of its processes change, especially its decision-making processes. Decisions also become data driven, that is, they include an analytic component based on the processing of internal or external data. This may be data stored in data lakes (which are increasingly used by companies for data storage), or in a database, whether structured or unstructured.

Similarly, the analysis can be particularly complex, in the field of data science, using techniques such as clustering, classification, logistic regression, decision trees and so on. But it can also be visual analysis techniques or simply inferential and descriptive statistics.

A decision process becomes data driven when it incorporates an analytic dimension. All these new roles and professions created around data seem to have the common mission of analyzing data in order to help decision-making whatever it may be.

From the decision-maker's perspective, the significance of big data lies in its ability to provide information and knowledge of value, upon which to base decisions. It allows decision makers to capitalize on the resulting opportunities from the wealth of historical and real-time data generated through supply chains, production processes, customer behaviors, etc. [FRA 08].

The analysis of big data involves multiple distinct phases that include data acquisition and recording, information extraction and cleaning, integration, aggregation and representation, query processing,

data modeling and analysis, and interpretation. Each of these phases introduces challenges.

2.5. Things to take into account

2.5.1. *Data complexity*

In early 1980s, while the subject of data began to gain importance, Hal B. Becker published *Can users really absorb data at today's rates? Tomorrow's?* The density of information at the time of Gutenberg was approximately 500 characters per square inch (worth approximately 2.54 cm). In 2000, it is anticipated that this capacity will be "1.25 × 1011" bytes per square inch.

The vision of Hal B. Becker has largely concretized since 2011, as data were recorded in zettabytes, 200 times more in a single year than what had been measured previously. Due to the advent of IT, modern datasets are evolving not only in term of size or volume or even diversity, but also in term of "complexity".

The amount of data that is traveling across the internet today is characterized by its complexity. Data can come from a variety of sources (typically both internal and external to an organization) and in a variety of types. With the explosion of sensors and smart devices as well as social networking, data in a company are more complex because they include not only structured traditional relational data, but also semistructured and unstructured data.

The Web enters a new phase of its existence as the largest and most dynamic reference point for data in the world. Data come from multiple sources, which makes it difficult to collect, transform and analyze. In this way, the term "variety" involves several different issues.

First of all, data – especially in an industrial environment – can be presented in several different ways, such as texts, functions, curves, images and graphs, or a combination of these elements. On the other

hand, these data show great variety, which often reflects the complexity of the studied phenomenon.

So data complexity is growing with the increase in its quantity, its velocity and diversification of its types and sources.

2.5.2. *Data quality: look out! Not all data are the right data*

Data are the indispensable raw material of one of the new century's most important activities. Be careful! In our analysis and predictions procedure, a lot of data are not yet "the right data". There is, therefore, underlying difficulty behind big data, since more data are not necessarily better data.

Since the implementation of the first data warehouse in the 1990s, the question of the quality of the data has been a major issue. In the United States, the theorem "Garbage In, Garbage Out" (GIGO) was immediately widespread.

So there is nothing new about this description: only data quality will help produce an event, a forecast or strategic information and define an action lever. The reconciliation of internal and external data has always been a challenge. It is possible to obtain better results by making better use of available data. When researchers encounter a set of data, they need to understand not only the limits of the available data, but also the limits of the questions that it can respond to, as well as the range of possible appropriate interpretations.

The ultimate goal is not only to collect, combine or analyze all data, but also to increase its value and efficiency. This means that we must evolve from the term "Big data" to "Smart data", since the effectiveness of companies' strategies now depends on the quality of data.

Indeed, companies must not rely on the size of their data – it is not useful unless it is applied in an intelligent manner. Therefore, the volume of data is of little importance, since internal data must be combined with external data in order for a company to obtain the most out of their data.

To make the most out of big data, the idea is not limited to the "simple" technical issues of collection, storage and processing speed. Big data uses require rethinking the process of collecting and processing, and the way data is managed. It is the "analysis" that will be applied to data that will justify big data, not the collection of data itself.

2.5.3. *What else?…Data security*

While big data regularly becomes an important issue of research and has been used everywhere in many industries, big data security has become an increasing concern. Nevertheless, there is a noticeable challenge between big data security and the uses of big data.

Big data and the IoT offer substantial development prospects for individuals and businesses. With the advent of big data comes the risk of greater security breaches as data volumes increase. Today's environment threatens the 3Vs of big data. Each of these is increasing at an astounding rate and require a shift in how security vendors manage threats.

The diversity and high volume of data sources, formats and data flows combined with the streaming nature of data acquisition create unique security risks. So, security issues are magnified by the 3Vs of big data. The use of large-scale cloud infrastructures with diversity of software platforms, spread across large networks of computers, also increases the attack surface of the entire system.

Therefore, traditional security mechanisms are inadequate. The variety, velocity and volume and the other additional Vs of big data amplify not only analytics tools but also security management challenges.

Data security not only involves the encryption of data, but also ensures that appropriate policies are enforced for data sharing. Security aims to preserve digital systems against malicious actions which target confidentiality, integrity or the availability of the system itself. An effective cyber security solution must deploy technical and organizational countermeasures while staying up to date in response to changing threats.

However, these developments imply that all actors have confidence in the systems and technological networks that underpin the digital revolution. As a condition of confidence, security is a major challenge for companies; the advent of big data, the cloud and mobility in an increasingly connected environment makes it even harder.

These new security challenges have led to increased security intelligence, which in turn requires big data and big data analytics.

2.6. Big data and businesses

2.6.1. *Opportunities*

Big data is somewhat unusual in that it was broadly accepted in the commercial and public space before the academic discourse had time to catch up [GAN 15]. This may explain why most of the literature on big data has increased during the last few years. Big data is not new, but the rapid rate of adoption in recent times may make it appear so [THE 17].

Big data is a natural crop of the advanced digital artifacts and their applications. Sensors, mobiles and social media networks are examples of modern digital technologies that have permeated our daily lives. A large amount of digital data is being generated every day.

These data are not only voluminous; they are also continuous, streaming, real-time, dynamic and volatile [SED 17]. What really is subsumed under big data is qualified under a few main characteristics:

1) Big data is primarily network-generated on a large-scale volume by variety and velocity and comprises large amounts of information at the enterprise and public level, in the categories of terabytes (10^{13} bytes), zettabytes (10^{21}) and beyond of online data. The trend is part of an increasingly popular environment: the proliferation of web pages, image and video applications, social networks, mobile devices, apps, sensors, and so on generates, according to IBM, more than 2.5 quintillion bytes per day to the extent that 90% of the world's data has been created over the few past years.

2) Big data consists of a variety and diversity of data types and formats, many of them are dynamic, unstructured or semistructured and hard to handle by conventional statistical methods.

3) Big data is generated by disparate sources such as interactive applications through the IoT from wireless devices, sensors and streaming communications generated by machine-to-machine (M2M) interactions.

Big data is revolutionizing how intelligence is stored and informative analysis can be drawn. The advent of the hyperconnected digital economy, powered by the IoT, is creating a new economy where data is the new commodity.

As a result, data have passed from being a modest and oft-discarded by-product of firms' operations to become an active resource with the potential to increase a firm's performance and economic growth. The literature indicates that big data can unlock plenty of new opportunities, and deliver operational and financial value [OHL 13, MOR 15, FOS 17, MCK 16].

The opportunity that big data presents for enterprises by tapping into varieties and volumes of data can bring benefits by informing plans and decision-making, discovering new chances for optimization and delivering breakthrough innovations. For that reason, companies are devoting their resources and efforts to gain greater results by leveraging big data.

The Internet has triggered a boom in information research. Companies are flooded by the wealth of data that results from simple Internet browsing. Pioneers in this field have led the way: Google, Facebook, Amazon and others. Data are at the heart of their business models.

The American company "Harrah's" has made progress in sales of 8–10% by analyzing customer segmentation data, while Amazon stated that 30% of its turnover came from its engine's analytical recommendations [MCK 11, MCK 13].

Such examples, and many others, share common principles: extreme digitalization of their process leads to extensive use of data to experiment with new business models, beyond their original boundaries. The exploration of large amounts of data enables the launch of new products and services, new processes and even new business models.

The "data revolution" has created new business models and investment strategies allowing companies to monetize their existing data and to create new big data solutions.

In addition, businesses can use open data to address civil society-focused issues, such as improved access to services or ratings of local service providers. Therefore, mangers and decision makers have to realize the potential of big data beyond a promising buzzword. They must pair vision with a clear profit model.

In other words, they are forced to purchase pertinent information to develop high added value strategies that allows them to succeed in the face of incessant changes in their business. Many examples (Google, Twitter, Facebook, Netflix, etc.) reveal that several opportunities are available to utilize big data; however, there are still many issues and challenges to be addressed to achieve better utilization of this technology.

2.6.2. *Challenges*

In order to harvest value from big data, companies have to address and overcome some challenges linked to:

1) *Big data dimension:* Big data gets global attention and can be best described using the 3Vs. Each dimension presents both challenges and opportunities for data management to advance decision making. The 3Vs provide a challenge associated with working with big data. The volume emphasizes problems with the storage, memory and capacity of a computing system and requires access to a computing cloud. Velocity stresses the rate at which data can be absorbed and meaningful answers produced, while the variety makes

it difficult to develop algorithms and tools that can address the large variety of input data [SED 17].

2) *Technological context:* There are many challenges of using and implementing big data. Manyika *et al*. [MAN 11] indicate that a key obstacle is the consistency of internal and external databases, implying that there is challenge in integrating and standardizing data of contrasting formats to enable valuable information flows. So, one of the main issues is the incompatible IT infrastructures and data architectures. IT systems and software should be able to store, analyze and derive useful information from available data (structured, semistructured and unstructured). The most successful companies understand the limitations of the technology behind their big data operations and recognize the importance of combining analysis with a sound understanding of the context, a good intuition for the industry and a critical attitude toward insights derived from data.

3) *Managerial context:* In the big data universe, companies seek to unlock the potential of data in order to generate value. The keystone of big data exploitation is to leverage the existing datasets to create new information, enriching the business value chain [SED 17]. The major challenge to overcome is management's lack of understanding of the potential value big data can bring to companies [MOR 15]. The goal is to manage the increasing amount of data and information, and to ensure its usage and flow across the companies. Data are required to be managed in different steps and most of all analyzed [KUD 14] for organizations to gain knowledge and value.

The challenges include not just the previous contexts, but also other issues related to scalability, heterogeneity, quality, timeliness, security and privacy.

– *Heterogeneity*: Data can be both structured and unstructured. They are highly dynamic and do not have a particular format. It may exist in the form of e-mail attachments, images, pdf documents, medical records, graphics, video, audio, etc. which cannot be stored in row/column format as structured data. Transforming these data to structured format for later analysis is a major challenge in big data analytics. However, machine analysis algorithms expect homogeneous

data and cannot understand nuance. As a result, data must be carefully structured as a first step in data analysis.

– *Scale*: Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore's law, to provide us with the resources needed to cope with increasing volumes of data. The difficulties of Big Data analysis derive from its large scale as well as the presence of mixed data based on different patterns or rules (heterogeneous mixture data) in the collected and stored data. Especially, in the case of complicated heterogeneous mixture data, the data have not only several patterns and rules but characteristically, the properties of the patterns vary greatly [FUJ 12].

– *Timeliness*: As the size of the datasets to be processed increases, they will take more time to analyze. In some situations, results of the analysis are required immediately. Therefore, businesses need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination [SED 17]. In big data, the realization time to information is critical to extract value from various data sources, including mobile devices, radio frequency identification, the web and a growing list of automated sensory technologies. Therefore, incorporating the time component requires a number of challenges. First of all, we have to make sure of the availability of the data. If the data emitted by the connected objects integrate the time component natively (via the concept of time series), their immediate storage and indexing require a particular framework.

– *Complexity*: Complexity measures the degree of inter-connectedness and interdependence in big data structures such that a small change in one or a few elements can yield very large or small changes that ripple across or cascade through the system and substantially affect its behavior, or no change at all [KAT 13]. Traditional software tools are not enough for managing the increasing volumes of data. Data analysis, organization, retrieval and modeling are also challenges due to the scalability and complexity of the data that need to be analyzed.

– *Quality*: Big data processing requires an investment in computing architecture to store, manage, analyze and visualize an enormous amount of data. It is the indispensable raw material of one of the new century's most important activities. But it is important to be prudent in our analysis and predictions because a lot of data are not yet "the right data". There is, therefore, underlying difficulty behind big data, since more data are not necessarily better data.

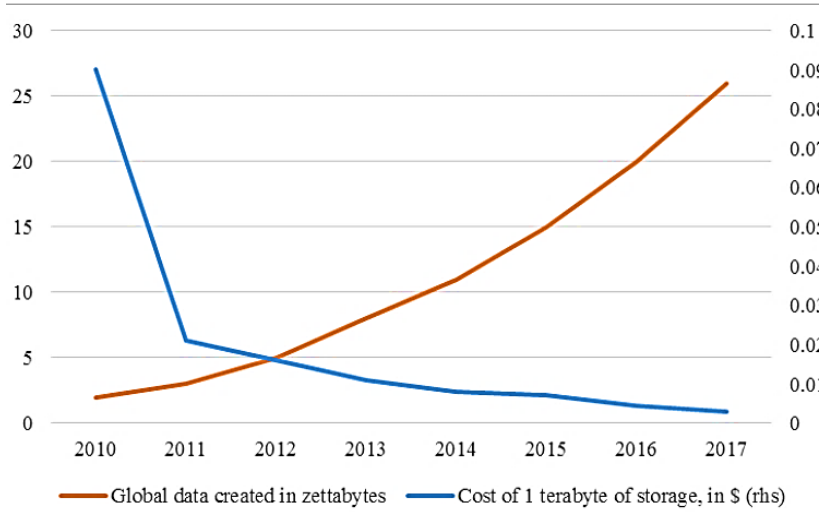
– *Security*: The vast majority of data comes from the many devices and machines reporting to each other and to those running them. From assembly lines at manufacturing plants to passenger jets in flight, millions of bytes of data are generated and then analyzed. Some of the captured data are personal information, and as such, both cutting-edge security and responsible stewardship models must be used to make sure this information is safe and correctly used.

– *Privacy*: The advance in big data analytics brought us tools to extract and correlates these data, which would make data violation much easier. That makes developing the big data applications a must without forgetting the needs of privacy principles and recommendations. The lawsuit following the Netflix challenge is a striking example where linking the provided data to IMDB movie reviews made if possible to identify certain users.

The nature of existing data (greatest dimension, diverse types, mass of data, structured and unstructured, etc.) does not authorize the use of most conventional statistical methods (tests, regression, classification, etc.). Indeed, these methods are not adapted to these specific conditions of application and in particular suffer from the scourge of dimension. These issues should be seriously considered in big data analysis and in the development of analytical procedures.

Similarly, there is cheaper storage, parsing (see Figure 2.3) and analysis of data through the availability of targeted databases, software and algorithms.

Companies are required to deal with these issues in order to seize the full potential of big data. So, data are not a power; it is their use that gives power, and the more one gives and exchanges data and information, the more one receives [MAR 01].



Source: Reinsel, Gantz and Rydning (2017); Klein (2017)

Figure 2.3. *Costs of storage and data availability (2009–2017) (source: [REI 17, KLE 17])*

2.7. Conclusions

Big data marks a major turning point in the use of data and is a powerful vehicle for growth and profitability. A comprehensive understanding of a company's data, its potential and the analytics methods can be a new vector for performance. Big data is a broad term generally referring to very large data collections that impose complications on analytics tools for harnessing and managing such.

Well-chosen and well-implemented methods for data collection and analysis are essential to better understand data. In another way, every piece of data tells a story and data analytics, in particular the statistical methods coupled with the development of IT tools, piece together that story to reveal the underlying message.

The next part of this book will provide us with the practical elements to understand this message by learning to ask good questions in order to be able to approach the problem as a "data scientist".

PART 2

Big Data Analytics: A Compilation of Advanced Analytics Techniques that Covers a Wide Range of Data

3

Building an Understanding of Big Data Analytics

> "If we are to go forward, we must go back and rediscover those precious values – that all reality hinges on moral foundations and that all reality has spiritual control".
>
> M.L. King Jr., "Rediscovering Lost Values" (1992)

3.1. Introduction

The rise of big data reflects the growing awareness of the "power" behind data and of the need to enhance gathering, exploitation, sharing and analyzing of data.

Analytics applications will ensure the proper exploitation of the proliferating volumes of data for a variety of business purposes, involving not only production of simple data-driven insights on operations, but also prediction of future trends and events.

As with all innovative areas, it is sometimes difficult to understand what is involved. That is why before going into the depths of the subject and talking only about the data analytics process, I suggest, in this chapter, that we first build an understanding of the development of analytics tools. Let's go!

Data Analytics and Big Data, First Edition. Soraya Sedkaoui.

3.2. Before breaking down the process… What is data analytics?

Before breaking down the process of data analytics in Chapter 5, and in order to understand big data analytics, it is necessary to look at what it is and under which category it falls. Many terms in business literature are often related to one another: "analytics", "business analytics" and "business intelligence" (BI).

Davenport and Harris [DAV 07] define analytics as: "the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions". An analytics team often uses their expertise in statistics, data mining, machine learning and visualization to answer questions and solve problems that management points out.

Analytics can also be defined as "a process that involves the use of statistical techniques (measures of central tendency, graphs and so on), information system software (data mining, sorting routines) and operations research methodologies (linear programming) to explore, visualize, discover and communicate patterns or trends in data" [SCH 14].

Business analytics begins with a dataset or commonly with a database. As databases grow, they need to be stored somewhere. Technologies, such as computer and data warehousing, store data. Database storage areas have become so large that a new term was devised to describe them.

Stubbs [STU 11] believes that business analytics goes beyond plain analytics, requiring a clear relevancy to business, a resulting insight that will be implementable, and performance and value measurement to ensure a successful business result.

Business analytics traditionally covers the technologies and application that companies use to collect mostly structured data from their internal legacy systems. These data are then analyzed and mined using statistical methods and well-established techniques classed as

data mining and data warehousing [CHE 12]. Generally, businesses perform two main types of analytics [DEL 13]:

– *descriptive:* which focuses on reporting on what happened in the past;

– *predictive:* which uses past data to try and predict future events.

Delen and Demirkan [DEL 13] noted that big data adds the ability to perform a third type of analytics, called perspective analytics, which combines data from the two previous types and uses real-time external data to recommend an action that must be taken within a certain time to achieve a desired outcome. So, there are many types of analytics (see Figure 3.1), and there is a need to organize these types to understand their uses.

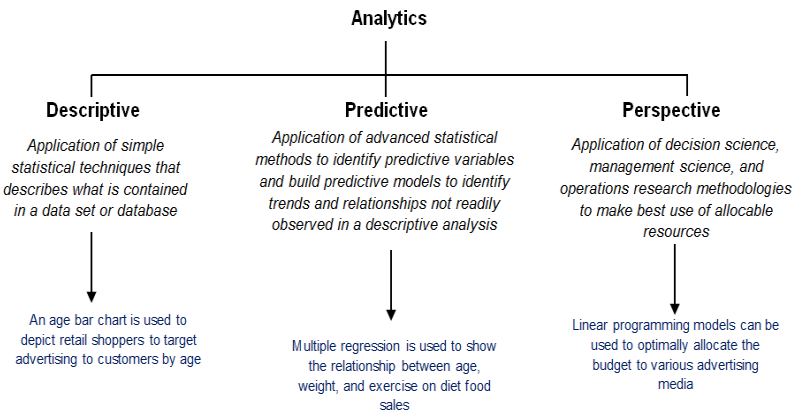


Figure 3.1. *Types of analytics*

Many companies from different sectors are looking for ways to exploit data in order to improve their operations and thus their skills which can monetarize these operations.

Some companies may only use descriptive analytics to provide information on decisions they face. Others may use a combination of types of analytics to glean insightful information needed to plan and make decisions.

Thus, the process of analytics can involve any one of these types, the major components of business analytics include all three used in combination to generate new, unique and valuable information that can aid business decision-making. In addition, the three types of analytics are applied sequentially (descriptive, then predictive, then prescriptive).

Therefore, business analytics can be defined as "a process beginning with business-related data collection and consisting of sequential application of descriptive, predictive and prescriptive major analytic components, the outcome of which supports and demonstrates business decision-making and organizational performance" [SCH 14].

Data processing and analysis, in the present day, are brought together under the notion of "Business Intelligence", due especially to computers' increased processing capabilities. According to Chen *et al*. [CHE 12], the term BI became popular in the 1990s, with the term "business analytics" added in the late 2000s to show the importance of analytical capabilities. Analytics has emerged as a catch-all term for a variety of different BI and application-related initiatives.

For some, it is the process of analyzing information from a particular domain, such as website analytics. For others, it is applying the breadth of BI capabilities to a specific content area (for example sales, service, supply chain etc.).

Data analytics is a process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. It focuses on knowledge discovery for predictive and descriptive purposes to discover new ideas or to confirm existing ideas.

It can be seen from the above definition that data analysis is a primordial step in the process of knowledge discovery in databases (KDD). This step involves the application of specific algorithms to extract patterns (models) from data.

The additional steps are data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge and proper

interpretation of the results of mining [MIT 02]. Powerful analytics tools can then be used to process the information gathered in large sets of structured and unstructured data.

3.3. Before and after big data analytics

Before the era of IT tools, company data was mainly in the form of handwritten paper records, which were not easily accessible. More recently, with advanced technology, larger amounts of data can be collected, stored and reused. And now, a new IT term is coined, i.e. the Internet of Things (IoT), in which everything is connected.

Therefore, this expands the amount of data, and consequently increases the importance of "Data Analytics".

Data analysis came in the 20th Century when the information age really began. Zhang mentioned in his book *Data Analytics* published in 2017 that the first real data processing machine came during the Second World War. But, the advent of the Internet sparked the true revolution in data analysis.

Davenport [DAV 14] states that company managers have been familiar with using traditional data analysis to support decisions since 1970. Vasarhelyi *et al*. [VAS 15] state that the traditional accounting data in companies were enterprise resource planning (ERP) data, which was acquired manually in transactions.

However, the importance of data analysis started in the late 1960s when researchers begin to speak about databases as repositories of data. Codd [COD 70] and his research group at IBM labs applied some mathematical principles and predicate logic to the field of data modeling. Since then, databases and their evolutions have been used as a source of information to query and manipulate data.

In 1974, still at IBM labs, the first language for databases was developed. SEQUEL (Structured English Query Language) [CHA 74], later called SQL for copyright issues, was the forerunner of all the query languages, becoming the standard for relational databases.

In the 1970s and 1980s, computers could process information, but they were too large and too costly. Only large firms could hope to analyze data with them. Codd was the first to work on data organization by designing database management systems, in particular of relational databases.

Since the 1980s, relational management systems have therefore taken precedence over other systems for the needs of all types of data, first for business and academic systems, then with independent developers for free initiatives or personal use, such as software, websites, etc. Even for small needs, embedded or local systems like SQLite (http://www.sqlite.org/) are widely used.

Quickly, a different need arose. The relational model is efficient for a purely transactional use, that which is called "OLTP" (Online Transactional Processing).

A management database, for example, used in ERP, has permanent activity updates and reduced result sets readings. We query the table of filtered invoices for a customer, which returns a dozen lines; we request the table of payments in order to verify that this customer is solvent; if so, we add an invoice with lines of invoices, and for each product added, it decrements its stock in the product table.

All these operations have limited scope in tables whose cardinality (the number of lines) can be otherwise important. But, because of a good data modeling, each of these operations is optimized. But what about statistical needs? How do we respond to requests for dashboards, historical analysis or even prediction?

In an ERP, there is a complete analysis of sales trends by product category, branch, department, month and by customer types, calculating developments to determine which product categories are changing, in which region and for which customer, etc.

In this kind of query, or what we call online analytical processing ("OLAP"), which has to cover a large part of the data to calculate

aggregates, the relational model and the query optimizers of the databases cannot respond satisfactorily to the need.

The OLAP model was created due to increased aggregated and historical data storage and global query requirements on these large volumes for analytical purposes. This is called BI. This model, which has also been formalized by Codd, prefigures the big data phenomenon.

In recent years, with the advent of Web 2.0 and the semantic web era, data analysis has become very important, replacing the traditional storing systems in many applications. They now represent the new technology for knowledge representation, data storage and information sharing.

3.4. Traditional versus advanced analytics: What is the difference?

Big data is often too large for data analysts to view and process on hand. The need for more advanced visualization techniques, capabilities to find patterns in complex data and modeling capabilities have increased along with the introduction of big data [SCH 14, IBM 12].

So, the key question is what is the difference between the two?

	Traditional analytics	Big data analytics
Technology	Analytics is done on batch jobs containing aggregated data, which are historical data rather than real-time data. Relational databases, data warehouses, dashboards.	A stack of tools that enables us to build a framework that makes it possible to extract useful features from a large dataset.
Skills	Basic knowledge of reporting and analysis tools.	Advanced: analytical, mathematical and statistical knowledge required to develop new models (data scientist).

	Traditional analytics	Big data analytics
Processing	Is appropriate for analysis of data containing information that will answer to information gaps that are known.	A lot of parallel processing often strains supporting systems. Can be truly ground-breaking for organization's as previously completely unknown gaps of information can be revealed randomly rather than just providing information about what is known.

Table 3.1. *Traditional versus advanced analytics*

Table 3.1 can help us to determine what is changing with big data analytics. But, to further clarify this difference, let us call out the example given by Krittika Banerjee (Research Analyst at Aspire Systems). Krittika invites us to imagine a quadrant, where the left of the x-axis refers to the "Questions we are asking", and the right presents "Questions we are not asking".

Similarly, the top of its y-axis represents "What we have sought", and the bottom represents, "What we haven't sought".

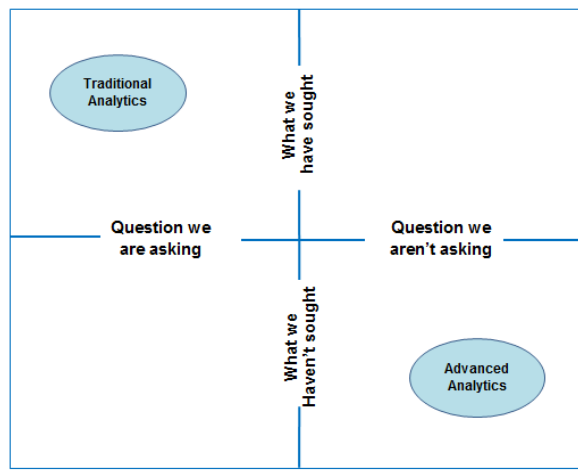


Figure 3.2. *The difference from another point of view (source: http://blog.aspiresys.com/digital/big-data-analytics/traditional-bi-vs-advanced-analytics/)*

The upper-left area of the quadrant is where traditional analytics would reside, while advanced analytics would be in the lower right section. The difference between the areas of the quadrant represents the distinction between these two approaches.

To incorporate our own approach in this book, we preferred to redesign the hierarchical model (or the knowledge pyramid) to better answer the question and understand the difference. This approach is discussed in Figure 3.3.

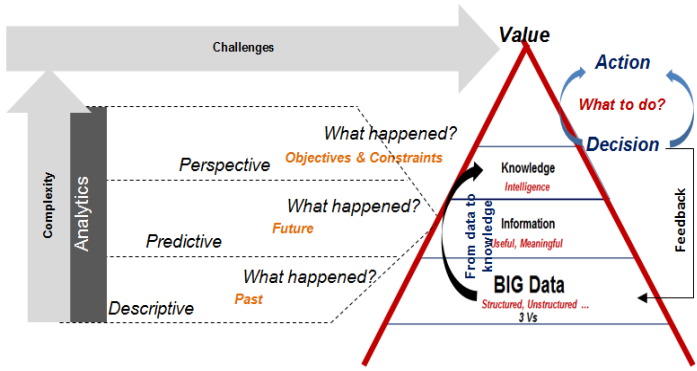


Figure 3.3. *Big data analytics: the road for knowledge*

Traditional analytics (descriptive) provides a general summary of data while advanced analytics deliver deeper data knowledge. Traditional analytics mines past data to report, visualize and understand what has already happened. While, modern analytics leverages past data to understand why something happened. Or to predict what will happen in the future across various scenarios. Advanced analytics, represented as prescriptive analytics in Figure 3.3, determines which decision and/or action will produce the most effective result against a specific set of objectives and constraints.

New analytics approaches in the big data age combine predictive and prescriptive analytics to predict what will happen and how to make it happen. Analytics' uses and applications improve the efficiency of decision-making processes and generate value.

The difficulty of transforming big data into value or knowledge is related to its complexity, which is growing with the increase in its quantity and velocity and diversification of its types and sources [SED 17]. Leveraging leading tools and techniques helps to manage and extract relevant data from big data. Advanced analytics can range from historical reporting to real-time decision support for organizations based on future predictions.

3.5. Advanced analytics: new paradigm

During the time of applying methods of statistical inference and statistical decisions some 70 years ago, information derived from data collection was considered costly. Models were built where information was linked to payoff relevance of a decision-making criterion (utility or payoff function), therefore statistical information was handled to satisfy these criteria.

Now as masses of data are produced at relatively low costs, all these data could be quickly aggregated. Statisticians have coined a term, "value of perfect information", which is set up to integrate data points, collection and analysis through statistical inferential models, i.e. exploratory data analysis or through statistical decision models [PIE 15]. However, achieving this goal is quite challenging all the data must be gathered for perfect information.

In traditional statistics, there are limited amounts of data, and it must obtain as much information as possible. In the big data age, there is a limited amount of computational power, and companies need to make the best decision.

The challenge in the analytical setting is that the analysis of subsets of data may present different analytical properties than the overall dataset.

For example, confidence intervals based on subsets of data will generally be wider than confidence intervals based on the original data; thus, care must be taken that the overall divide-and-conquer procedure yields a correctly calibrated interval [JOR 13].

The continuous increase in measurement capabilities and development of new data uses make it increasingly necessary to have analytical tools to summarize and extract information from data.

However, the nature of modern data (greatest dimension, diverse types, mass of data) does not authorize the use of most conventional statistical methods (tests, regression, classification). Indeed, these methods are not adapted to these specific conditions of application and in particular suffer from the scourge of dimension.

When the dimensions of problems (n, p) are reasonable and those of model assumptions (linearity) and distributions are checked, or in other words, usually, when the sample or residue are assumed to follow laws putting themselves in the form of an exponential family (Gaussian, binomial, etc.), modeling statistical techniques drawn from the general linear model are optimal (maximum likelihood) and, especially in the case of small samples, it seems difficult to do much better.

As soon as the distributional assumptions are not verified, the assumed relationships between variables or variable to be modeled are not linear or when the volume of data is too large, other methods compete with the traditional statistical approach.

Consider a simple example to explain a quantitative variable Y through a set $\{X1, \ldots Xp\}$ of quantitative variables:

$$Y = f(X1, \ldots Xp) + \varepsilon$$
$$(yi, xi),\ i = 1,\ldots, n$$

If the function is assumed to be linear and p is small, of the order of 10, the problem is well known and widely discussed in the literature. In the case where the function f is not exactly linear and n is large, it is possible to accurately estimate a larger number of parameters and therefore to envisage more sophisticated models.

If we consider the usual Gaussian model, even the simplest case of a polynomial model quickly becomes problematic. Indeed, when the function is linear, take $p = 10$, the model selection procedure is facing

a group of 2^{10} possible models and only sophisticated algorithms are able to cope.

However, considering an estimation of f, a simple polynomial of the second or third degree, with all its interactions, leads us to consider a large number of parameters and thus, by combinatorial explosion, an astronomical number of possible models. Other methods must then be considered taking into account, necessarily, the algorithmic computational complexity.

This explains the involvement of another discipline, i.e. computing. The concern of computability outweighs the mathematical definition of the problem, which optimizes a criterion adjustment of the function f over a set of more or less rich solutions.

These methods have often been developed in another disciplinary environment: computers, artificial intelligence, K-means, neural networks, decision trees, support vector machines, etc. become credible alternatives since the number of observations is sufficient or the number of variables is very important.

3.6. New statistical and computational paradigm within the big data context

Statistics is the traditional field that deals with the quantification, collection, analysis, interpretation and evaluation of data. The development of new statistical methods is an interdisciplinary field that draws on computer sciences, artificial intelligence, machine learning, visualization models and so on.

There are several methods that have recently been developed and are feasible for statistical inference of big data and workable on parallel machines, including the bag of little bootstraps, aggregated estimation equation and so on. Each method was developed to find and design tools that explicitly reveal tradeoffs relating complexity, risk and time.

Concerning statistical methods, the literature summarizes the change in two points [SED 17]:

1) *the new approaches are on the crossroads of IT tools and statistics:* this concerns machine learning, where algorithms generate, more or less alone models on large amounts of data;

2) *these methods are not new because machine learning dated from the 1960s:* this return to the center stage is due to the fact that these techniques work especially well on high amounts of information.

Big data poses new challenges to statisticians both in terms of theory and application. Some of the challenges include size, scalability of statistical computation methods, non-random data, assessing uncertainty, sampling, modeling relationships, mixture data, real-time analysis of streaming data, statistical analysis with multiple kinds of data, data quality and complexity, protecting, privacy and confidentiality, high dimensional data, etc.

As the volume of data grows, so do the requirements for more advanced data warehouses and dispersed cloud-based databases [KIM 11].

In cases of data analytics, we analyzed requirements regarding (1) data: types, structure, format and sources and (2) data processing: operations, performance and conditions.

The systematic application of data as a key driver for improving the robustness of decision making is widely considered a valuable, even necessary, practice for businesses. McAfee and Brynjolfsson [MCA 12] suggest that firms that consider themselves “data-driven” achieve consistently higher performance on several financial and operational measures, compared to those that do not.

It is focused on the development of methodologies and techniques that “make sense” out of data. It would require tailored analytical methods and data quality control to superimpose on large data streams to make sense of the data and use them for statistical inference and decisions [SED 17]. More often than not, good theoretical insights and models of the subject discipline would be useful in identifying the “payoff relevance” of data for predictive purposes [HAR 14].

The notion of making sense of big data has been expressed in many different ways, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology and data pattern processing.

In this book, big data analytics, or advanced analytics, is considered as an umbrella concept for the analysis of data with the explicit aim of generating value, in the form of efficient information, that aids the decision-makers in their process. This idea can be formalized by Van Barneveld *et al.*'s [VAN 12] definition:

> "Analytics is the process of developing actionable insight through discovery, modeling and analysis, and interpretation of data".

Where:

– The idea of *actionable insight* is applied to convey that the objective of analytics is to generate results that directly increase the understanding of those involved in the decision-making process [COO 12].

– *Discovery* refers to the problem definition and exploratory element of analytics; the identification, collection and management of relevant data for subsequent and/or concurrent analysis. This discovery stage integrates Cooper's [COO 12] emphasis of a problem definition with what Labrinidis and Jagadish [LAB 12] conceptualize as data management that includes the following:

- *problem definition*: identify what data to collect, and to subsequently begin acquiring it. But, the volume of data manipulated by some companies, especially those related to the Internet, increases considerably. The increasing computerization of all types of processing implies an exponential multiplication of this volume of data that now counts in PB, EB and ZB. Chen *et al.* [CHE 12] highlight the multitude of techniques that allow organizations to tap into text, web, social networks and sensors, all of which enable the acquisition and monitoring of real-time metrics, feedback, and progress.

- *Data collection*: The collection and combination of semistructured and unstructured data requires specific technologies, which also have to account for data volume and complexity.

- *Data management*: Data management involves the storage, cleaning and processing of the data.

– *Modeling and analysis* are concerned with applying statistical models or other forms of analysis against real-world or simulated data. The middle stage of this categorization involves making sense of the acquired data to uncover patterns and to evaluate the resulting conclusions [TOM 16].

– *Interpretation* involves making sense of the analysis results and subsequently conveying that information in the most comprehensible form onwards to the relevant parties. On the other hand, making sense of different types of data and generating value from it result in some form of finding.

But, it is necessary to point out that there are two computational barriers for big data analysis: the first concerns the data that can be too big to hold in a computer's memory, while the second is related to the computing task that can take too long to generate the results. These barriers can be approached either with newly developed statistical methodologies and/or computational methodologies [WAN 15].

From an IT point of view, knowledge of Hadoop is highly desirable. It allows for the creation of distributed applications and is "scalable" on thousands of nodes to manage petabytes of data. The principle is to split and parallelize (distribution) data batch tasks to linearly reduce the computation time (scalable) depending on the number of nodes. Hadoop becomes the mining web reference tool for e-commerce.

From a statistical point of view, the new challenge is both the functional representation of bases of construction and relevant models to address and take into account the complex data structures: geolocation on graphs, real-time signal, 3D images, sequences, etc.

Every problem, especially industrial ones, requires a specific approach after a search for a conventional engineering development. In the case of data streams, the decision support becomes adaptive or sequential. The computational tools that often are associated with the analysis of big data can also help scholars who are designing experiments or making causal inferences from observational data.

3.7. Conclusions

Faced with the volume and the diversification of data available today, it is essential to develop techniques to make best use of all of these stocks in order to extract the maximum amount of information. Indeed, a shift is also expected to be made in thinking; this could be about infrastructure of data but also about business intelligence and analytics.

Applying big data analytics is not about only knowing R or Python programming language... It is mainly about knowing why and how to apply the different technical tools. That is what will be illustrated in Chapters 4 and 5.

4

Why Data Analytics and When Can We Use It?

> "The alchemists in their search for gold discovered many other things of greater value".
>
> Arthur Schopenhauer, German Philosopher

4.1. Introduction

The increase in data produced by companies, individuals, scientists and public officials, coupled with the development of IT tools, offers new analytical perspectives. Analysis of big data requires an investment in computing architecture to store, manage, analyze and visualize an enormous amount of data.

If Facebook, Google, Twitter, LinkedIn, Amazon, Apple, Netflix, Nike and many other data-driven business models exist, it is because of the advantages generated by big data and analytics. But, what can analytics tools be used for? How have these models been able to gain such a profile? And what are the challenges that big data analytics is dealing with? This is what will be discussed in this chapter.

Data Analytics and Big Data, First Edition. Soraya Sedkaoui.

4.2. Understanding the changes in context

Increasingly, we discuss the benefits of the data analysis from Twitter, Google, Facebook and any other space, in which more and more people are leaving digital traces and filing information that may be exploitable and exploited.

Every second, visitors interact with interconnected objects and leave behind a tremendous amount of data that companies can then use to create tailor-made experiences. Faced with such a challenge, both make sure that the technologies used are able to correctly handle this volume of data.

Big data and the use of data analytics are being adopted more frequently, especially in companies that are looking for new methods to develop smarter capabilities and tackle challenges in dynamic processes.

Big data and profit are closely intertwined. The correlation identified by Google between a handful of search terms and the flu is the result of testing 450 million mathematical models. The United Nations has developed a program anticipating epidemics and economic downturns through key words exchanged on Twitter.

Danish company "Vestas Wind Systems", one of the largest manufacturers of wind turbines in the world, uses "IBM Big Data Analytics" and "IBM Systems" solutions to decide the location of wind turbines by crossing varied data in a matter of hours (meteorological and geo-spatial data, satellite images, etc.).

The case of "BlaBlaCar" in this area illustrates the power that analytics brings to the management of a carpooling service. The startup has the advantage of conducting change management to implement the new analytics solutions. Beyond A/B testing, the work on big data consists of analyzing user behaviors, optimizing the interface, etc.

Big data analytics has become an essential requisite to run most businesses. Integrating big data analytics can generate many advantages for the companies, such as:

– *supporting decision*: as mentioned before, companies can make use of the vast amount of data relevant to their particular business. Therefore, they would need to filter the data according to their specific needs and derive meaning from the data that fits them best. This will not only widen their understanding of their own domain but will also facilitate better decision-making, which in turn will improve operational efficiencies;

– *cost reduction*: it has been found that big data can be extremely instrumental in augmenting the existing architectures of companies. Additionally, when more accurate decisions are taken, the possibility of incurring losses also gets alleviated. Therefore, with the correct use of analytics, businesses can be successful in cutting down their operational costs, which is typically one of the biggest challenges;

– *customer insights*: the growth of any company depends on how to take into account the preferences, likes, tastes, etc. of their customers in the design of their products and services. Big data analytics can help companies to gain access to the required and relevant information. For example, social media presents a great tool to acquire and assimilate enormous volumes of customer insights and can be used effectively to collect data for this purpose;

– *open data uses*: over the last year, there has been an increase in the perceived use of open data to build new products and services. Open data, in addition to its economic potential and the creation of new activities it entails, also falls within a domain of philosophy or ethics. It belongs to individuals and can be used to encrypt their behavior. The culture of this phenomenon builds on the availability of data for a communication orientation.

Combined with advanced analysis methods, new explanations can be provided for several phenomena.

The analyzed data allow companies to obtain strategic advantages by taking into account a greater number of data, by improving existing ones (optimization and cost reduction, productivity gains, etc.), by

creating new ones, more targeted products, or to improve the customer experience. Indeed, better knowledge of new needs is clearly an asset.

Because the data state a lot about customer preferences, they represent business and marketing issues for the company. Data mining, for example, is a technique that can extract and analyze big data to extract relevant information. Simply put, data mining software processes the huge amount of data to highlight trends, models and correlations. For example, if your company sells air conditioners, you may have noticed that sales increase during the summer.

Data mining makes it possible to be much more precise and to highlight that the sales of these kinds of products increase a few days after a heat wave, since the average maximum temperature of these days exceeds 28 degrees. This makes it possible to accurately predict when the demand will increase (or not), to accordingly adapt the rate of production and the supply chain or to launch an advertising campaign at the right time.

The models and correlations established by data mining not only make it possible to understand the present, but also to anticipate behaviors. In this, it serves as a basis for machine learning.

Industrial maintenance, the customization of offers, energy efficiency, preventive medicine and the autonomous car are all examples of the different applications of big data analytics. The profound transformations engendered by large-scale data processing capacity are able to redefine the boundaries between different sectors and companies that will seize the opportunity.

Cities and even entire countries also benefit from data analysis. Thus, real-time resource management is now possible. For example, the Digital Delta project, launched by the Netherlands in collaboration with IBM, aims to build a platform to support the distribution of water. It should eventually lead to a 15% decrease in the cost of managing blue gold. The Smart Cities market is estimated to be at more than $ 100 billion by 2030, globally.

4.3. When real time makes the difference

The challenge is not only to collect the data, but also to exploit, process and analyze them better, in such a way that allows businesses to generate knowledge in order to upgrade the process of decision-making and achieve higher performance.

Beyond the advent of ICT and increased data production, dissemination and processing speeds, another element has recently become critically important: "time". The importance of time carries with it a notion of information circulation speed [SED 16].

So, big data adds an unprecedented dimension: exploiting the profusion of huge volumes of data with the finest level of detail and often the shortest lifetime (instantaneity). And it will be more performant if it becomes possible to analyze the data in real time.

Traditional data processing architectures know how to collect data before transforming and then analyzing them. These three operations are globally performed one after another.

Today a number of elements are combined to define an integrated system of "big data injection in real time", which is about to transform business operations. Real-time analytics was growing at a rapid pace and is now poised to reach an inflection point.

This evolution could not come at a more opportune moment. Indeed, according to IDC, we are swept into a digital whirlwind that is growing by 40% per year. This growth is fueled not only by the presence of more and more businesses and individuals on the internet, but also by the rapid development of the IoT. This digital world doubles in size every 2 years: by 2020, predicts IDC, it will reach 44 zettabytes (ZB) (44 billion GB).

In descriptive, predictive and prescriptive analytics, one exports a set of historical data for batch analysis. In real-time analytics, one analyzes and visualizes data in real time.

For example, now, with the power of real-time data processing, a health service provider can continually monitor patients at risk. By combining the real-time data recorded by several connected object to monitor medical symptoms with information in medical records, analysis tools can alert health professionals if proactive action is urgent for the patient.

4.4. What should data analytics address?

The analysis of a larger amount of data in real time is likely to improve and accelerate decisions in multiple sectors, from finance to health, both including research. The considerable increase in the volume and diversity of digital data generated, coupled with big data technologies, offers significant opportunities for value creation.

This value cannot be reduced to simply what we can solve or improve, but rather it knows what the new potential discoveries are that may arise from cross-exchanges and correlations. This leads us to say that new data processing tools are now necessary, as are methods capable of combining thousands of datasets.

It is the use of data that empowers decision-making. Being increasingly aware of the importance of data and information, companies are pressed to rethink the way to "manage", to enrich and to benefit from them. This causes two main challenges as follows:

1) big data contains invisible models, which must be viewed using tools and analytical techniques. The knowledge gained should be used at the right time in the right context and with the right approach;

2) capturing, managing, combining, securing and always taking advantage of a huge amount of data are much more complicated than the simple data storage problem.

As large datasets are currently available from a wealth of different sources, companies are looking to use these resources to promote innovation, customer loyalty and increase operational efficiency. At the same time, they are contested for their end use, which requires a

greater capacity to collect, analyze and manage the growing amount of data but also ensure its security.

This highlights that it is not merely the existence of large amounts of data that is creating new security challenges. Data exploration and analysis turned into a difficult problem in many sectors in the case of big data.

Let us think about big data in network cybersecurity, an important problem. Governments, corporations, financial institutions, hospitals and other business collect, process and store confidential information on computers and transmit that data across networks or other computers.

With large and complex data, computation became difficult to be handled by the traditional data processing applications and triggered the development of big data applications [MUH 13]. If big data are combined with predictive analytics, they produce a challenge for many industries. The combination results in the exploration of these four areas are as follows [INU 14]:

– calculate the risks on large portfolios;

– detect, prevent and re-audit financial fraud;

– improve delinquent collections;

– execute high-value marketing campaigns.

There are many technical challenges that must be addressed to realize the full potential of big data. Warren *et al*. [WAR 15] state that many companies are unable to apply big data techniques due to limiting factors, such as lack of data, irrelevant or untrustworthy data, or insufficient expertise. The main challenges associated with the development and deployment of big data analytics are as follows:

– *the heterogeneity of data streams*: dealing with semantic interoperability of diverse data streams requires techniques beyond the homogenization of data formats. Big data streams tend to be multimodal and heterogeneous in terms of their formats, semantics and velocities. Hence, data analytics typically expose variety and

veracity. Big data technologies provide the means for dealing with this heterogeneity in the scope of operationalized applications;

– *data quality:* the nature of data available can be classified as noisy and incomplete, which creates uncertainty in the scope of the data analytics process. Statistical and probabilistic approaches must therefore be employed in order to take into account the noisy nature of data. Also, data can be typically associated with varying reliability, which should be considered in the scope of their integration in the analytical approach;

– *the real-time nature of big datasets:* big data feature high velocities and for several applications must be processed nearly in real time. Hence, data analytics can greatly benefit from data streaming platforms, which are part of the big data ecosystem. IT, the Internet and several connected objects typically provide high-velocity data, which can be in several cases controlled by focusing only on changes in data patterns and reports, rather than dealing with all the observations that stem from connected objects;

– *the time and location dependencies of big data*: IoT data come with temporal and spatial information, which is directly associated with their business value in an analytics application context. Hence, data analytics methods must in several cases process data in a timely fashion and from proper locations. Cloud computing techniques (including edge computing architectures) can greatly facilitate timely processing of data from several locations in the case of large-scale deployments. Note also that the temporal dimensions of big data can serve as a basis for dynamically selecting and filtering streams toward analytics tools for certain timelines and locations;

– *privacy and security sensitivity*: big data are typically associated with stringent security requirements and privacy sensitivities, especially in the case of IoT applications that involve the collection and processing of personal data. Hence, advanced analytics need to be supported by privacy preservation techniques, such as the anonymization of personal data, as well as techniques for encrypted and secure data storage;

– *data bias*: as in the majority of data mining problems, big datasets can lead to biased processing and hence a thorough

understanding and scrutiny of both training and test datasets is required prior to their operationalized deployment. Note that the specification and deployment of IoT analytics systems entails techniques similar to those deployed in classical data mining problems, including the understanding and the preparation of data, the testing of the analytics techniques and ultimately the development and deployment of a system that yields the desired performance and efficiency.

These challenges are evident in the big data analytics system, which comprises a series of steps from data acquisition to analysis and visualization. Jagadish *et al*. [JAG 12] provide a comprehensive discussion of such challenges based on the notion of the data analysis pipeline:

– *data acquisition and recording*: it is critical to capture the context in which data have been generated, to be able to filter out non-relevant data and to compress data, to automatically generate metadata supporting rich data description and to track and record provenance;

– *information extraction and cleaning*: data may have to be transformed in order to extract information from it and express this information in a form that is suitable for analysis. Data may also be of poor quality and/or uncertain. Data cleaning and data quality verification are thus critical;

– *data integration, aggregation and representation*: data can be very heterogeneous and may have different metadata. Data integration, even in more conventional cases, requires huge human efforts. Novel approaches that can improve the automation of data integration are critical as manual approaches will not scale to what is required for big data. Also, different data aggregation and representation strategies may be needed for different data analysis tasks;

– *query processing and analysis*: methods suitable for big data need to be able to deal with noisy, dynamic, heterogeneous, untrustworthy data and data characterized by complex relations. However, despite these difficulties, big data, even if noisy and

uncertain, can be more valuable for identifying more reliable hidden patterns and knowledge compared to tiny samples of good data. Also the (often redundant) relationships existing among data can represent an opportunity to cross-check data and thus improve data trustworthiness. Supporting query processing and data analysis requires scalable mining algorithms and powerful computing infrastructures;

– *interpretation and visualization*: analysis results extracted from big data needs to be interpreted by decision-makers and this may require the users to be able to analyze the assumptions at each stage of data processing and possibly retrace the analysis. Rich provenance is critical in this respect.

Figure 4.1. *Big data pipeline (source: [AGR 11])*

This process is supported by cloud computing and computational tools, including data mining, statistical computing and scalable databases technology.

4.5. Analytics culture within companies

In all sectors, the uses of big data analytics testify to the effectiveness of this technology. Hospitals and healthcare practitioners are collecting information about patients as they can predict epidemics and design new treatments that can reduce waste and improve service delivery.

Another potential benefit of big data is that it can provide more regular and timely information on interesting patterns, such as early indicators of epidemics, economic upturns or downturns, e.g. Google's flu indicators despite its problems, unemployment or housing boom etc., because of the lower unit cost of acquiring big data

sources than the traditional direct data collection methods used by NSOs.

An excellent example of this is provided in [CHO 11] who also coined the term *nowcasting* to describe the process of predicting the present by harnessing information from Google Trends. In a blog to the Washington Post, Mui [MUI 14] argued that the currency of statistics afforded by big data – readily available as a by-product of other collections – and how they could be "mined" for interesting patterns are a promising benefit of big data over traditional data sources.

Company	Big data project
Google	MapReduce, BigTable
Amazon	Dynamo
Yahoo	Hadoop
Facebook	Cassandra, Hive
Twitter	Storm, FlockDB
LinkedIn	Kafka, SenseiDB, Voldemort
LiveJournal	Memcached

Table 4.1. *The leaders in the big data analytics field*

Given the culture and business model of these companies, it is not surprising that most of these projects are open source. The case of these companies in various fields illustrates the power that analytics brings for management of the proposed services and applications available.

Prediction approaches and analytics methods have strong roles in the creation of social and economic opportunity. But, they are not only

a brand that bears the large companies. Analytical practices also feature in every entrepreneur's toolkit.

Many successful entrepreneurs' experiences support analytics as a core capability of their startups. These include Sergey Brin and Larry Page of Google, Jeff Bezos of Amazon.com, Michael Bloomberg of Bloomberg LP and Reed Hastings of Netflix. They have seen the potential of using analytics not only to differentiate but also to innovate their business models.

The analysis of big data is not only a matter of solving computational problems, even if those working on big data come from the natural sciences or computational fields. Rather, expertly analyzing big data also requires thoughtful measurement [PAT 15], careful research design and the creative deployment of statistical techniques.

Indeed, massive datasets will require the full range of statistical methodology to be brought to bear in order for assertions of knowledge on the basis of massive data analysis to be reliable.

Following a period when the main issue is how to organize and structure databases? The question now is what to do? What kind of analysis must be applied, adopted and developed to value and support decision-making? In another words, how should analytical approaches be designed so as to be scalable computationally to the massive datasets?

Then, to capitalize on its potential, companies must put data analytics at the center of their strategy. What is truly necessary are excellent analytic and soft skills, a capacity to understand and manipulate large sets of data, and the capacity to interpret and apply the results.

But, they also need to establish clear guidelines for data integrity and security, as digital ecosystems can only function efficiently if all parties involved can trust in the security of their data and communication.

So, it is clear that there are many varieties of factors that have contributed to the growing use of big data analytics and its applications. Generally, these factors have also spurred adoption of "Artificial Intelligence" (AI) and "Machine learning" (ML) in the business context. AI and ML are being adopted widely in sectors such as oil, industry, health, manufacturing, marketing and many other areas. These include availability of computing power owing to faster processor speeds, lower hardware costs and better access to computing power via cloud services [INS 17].

In the age of the data revolution, analytics is empowering companies to take a pragmatic, evidence-based approach to all aspects of their business, including communications and marketing, operations, transportation and logistics, cyber security and risk management.

4.6. Big data analytics application: examples

The new analytical power is seen as an opportunity to invent and explore new methods, which are able to detect correlations between the quantities of available data. Cukier and Mayer-Schoenberger [CUK 13a, CUK 13b] see a paradigmatic change in the statistical handling of large data:

> "Using great volumes of information … require three profound changes in how we approach data. The first is to collect and use a lot of data rather than settle for small amounts or samples as statisticians have done for well over a century. The second is to shed our preference for highly curated and pristine data and accept messiness: in an increasing number of situations, a bit of inaccuracy can be tolerated, because the benefits of using vastly more data of variable quality outweigh the costs of using smaller amounts of very exact data. Third, in many instances, we will need to give up our quest to discover the cause of things, in return for accepting correlations. With big data, instead of trying to understand precisely why an engine breaks down or why a drug's side effect disappears, researchers can instead collect and analyze

> massive quantities of information about such events and everything that is associated with them, looking for patterns that might help predict future occurrences. Big data helps answer what, not why, and often that's good enough".

Manyika *et al*. [MAN 11] argued that:

> "… there are five broad ways in which using big data can create value. First, big data can unlock significant value by making information transparent and usable at much higher frequency. Second, as organizations create and store more transactional data in digital form, they can collect more accurate and detailed performance information on everything from product inventories to sick days, and therefore expose variability and boost performance. … Third, big data allows ever-narrower segmentation of customers and therefore much more precisely tailored products or services. Fourth, sophisticated analytics can substantially improve decision-making. Finally, big data can be used to improve the development of the next generation of products and services".

Big data analysis is essential when organizations want to engage in predictive analysis, natural language processing, image analysis or advanced statistical techniques such as discrete choice modeling and mathematical optimization, or even if they want to mash up unstructured content and analyze it with their BI.

Companies will be able to suggest data management for decision-making. The new analytical power is seen as an opportunity to invent and explore new methods, which are able to detect correlations between the quantities of available data. Large companies are increasingly using big data analytics to improve their business. Netflix and Uber are two other examples of successful use of the latest analytics technologies.

Millions of Netflix subscribers generate a lot of data. To analyze these data, Netflix uses analytical techniques to determine what users will enjoy watching. Netflix's recommendation engine, impressive insight, works through analytics. Netflix has been able to build predictive models to suggest series that will delight users.

Netflix has developed other techniques for enhancing relevant recommendations, which is based on keywords. Subsequently, users receive suggestions based on these keywords, corresponding to the productions they have most appreciated.

In 2015, the message sent by Netflix to its shareholders showed that the big data strategy was paying off. In the first quarter of 2015, 4.9 million new subscribers were registered, compared with four million in the same period in 2014. Similarly, 10 billion hours of content were broadcast. Because of intelligent use of big data analytics, the influence of Netflix continues to grow.

The Uber smartphone app connects passengers and drivers with a principle based on big data analytics (crowdsourcing). Anyone who is willing to offer their driver services can offer their help easily. During each trip, Uber collects and analyzes data to determine the extent of demand across geographic areas. This allows the company to allocate its resources efficiently.

Uber also analyzes the public transport networks in the cities where it operates. In this way, the company can focus primarily on underserved areas. In addition, Uber has developed algorithms to monitor real-time traffic conditions and travel times. As a result, prices can be adjusted as demand and travel times fluctuate. Then, drivers tend to drive when they are needed most.

This pricing method based on big data analytics is patented by Uber. It is called "surge pricing". This is an implementation of the "dynamic pricing" already commonly used by airline companies and hotel chains to adjust the price on demand in real time through predictive analysis.

Other companies, which are considered as leaders in this field, have developed their strategy based on big data analytics. The e-commerce giant Amazon recommends products to customers based on their browsing and purchasing habits. The "ad-tech" companies such as RocketFuel apply statistical and optimization techniques to determine which banner ads to display.

Thus, devices such as "Fitbit" used for recording and monitoring our physical activities, and their integration with other applications, allow individuals to obtain information on calories burned and food consumed. This allows a creation of new models that sell this information to insurance companies to better calculate risks.

Also, Chicago city uses an algorithm (based on partly secret data) that has identified the city's 400 people most likely to commit acts of violence with a rating of danger for each.

Crime data can also threaten the traditional insurance model, which distributes risk over wide areas, by providing insurers with a much more granular view of risk (potentially down to the individual level). This is why open data are so attractive to innovators and entrepreneurs. The consequences for insurance pricing and the premiums paid by individuals deemed to be high risk are significant.

Terapeak provides e-commerce businesses powerful tools to optimize listings, source inventory, evaluate sales and find products on leading e-commerce platforms such as Amazon, eBay and Ali Baba.

Also, Shell uses big data and industrial IoT to develop a "data-driven oil field" that brings down the cost of production, monitors equipment in real time, manages cyber risks and increases efficiency of transport, refinement and distribution.

Many other companies use data analytics. This is the case of NASA, Domino's Pizza or the NFL. Nest has used data analytics to create a smart thermostat that optimizes electricity consumption by monitoring temperature, resident presence and more.

A series of announcements, from the acquisition of Nest Labs by Google for $3.2 billion to Samsung Gear and health-related wearables to the development of Smart Home features into Apple's iOS, have made big data an increasingly tangible business opportunity.

4.7. Conclusions

The process of gathering, processing and interpreting information is not limited to defining ideas, but also consists of materializing them in order to ensure improved knowledge production that leads to value creation.

Big data analytics can be defined as a new discipline born from the mixture of statistics, computer sciences and business. It allows each company to optimize its operation and strategy. The data must be analyzed and reviewed to be able to add value, especially when the aim is to optimize methods and operations.

After identifying the importance and role of analytics, is it now time to move on to know how it works?

We are undoubtedly impatient to discover it. So, let us follow Chapter 5 in order to understand the data analytics process.

5

Data Analytics Process: There's Great Work Behind the Scenes

> "I never guess. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts".
>
> Sherlock Holmes, *A Study in Scarlett*
> Arthur Conan Doyle

5.1. Introduction

With the wide usage of computers and Internet, there has recently been a huge increase in publicly available data that can be analyzed. Analyzed data are no longer necessarily structured in the same way as in traditional analysis, but can now be text, images, multimedia content, digital traces, connected objects, etc.

If we have seen, in previous chapters, the importance of big data analysis (the why?), as with every major innovation, the biggest confusion lies in the exact scope (what?) and its implementation (how?).

In this chapter, we will discover what is the usual cycle when working with big data, and understand at what stage data analytics

Data Analytics and Big Data, First Edition. Soraya Sedkaoui.

intervenes. So, here is what you need to know to get your feet wet with big data analytics.

5.2. More data, more questions for better answers

Before tackling the subject of a data analytics process, some points deserve to be relieved.

5.2.1. *We can never say it enough: "there is no good wind for those who don't know where they are going"*

Two essential components are needed to question whether data analytics can or cannot add value: *data* and a *well-defined problem*. Determining what type of problem you are facing will allow you to correctly choose the technique that can be used.

Everything in big data analytics begins with a clear problem statement. The success of an analytics approach cannot be possible without the clarification of what you want to achieve. This is not just valid in the context of big data, but in all areas. We must clearly define what we want before undertaking anything.

This means knowing what we are trying to achieve, what is needed, and why and what level of accuracy is acceptable and actionable.

What should be done in this phase is to explore all possible paths to recover the data in order to identify all of the variables that affect, directly or indirectly, the phenomenon that interests us. In other words, how do we make new opportunities from these data? Which data should we select for the analysis? And how do we efficiently apply analytical techniques in order to generate value?

What new insights can I expect? Where does the greatest global potential of big data lie? How will these insights help me? Having the ability to think critically allows you to understand that big data opportunities are not in the volume of data, but in the digital transformation of your business processes.

5.2.2. *Understanding the basics: identify what we already know and what we have yet to find out*

As we have already said, data are literally the nerves of the era of big data. The data are mostly available, but often scattered in several computer tools. An important procedure is to understand the data that will be collected and then analyzed. The idea is that the better your understanding of your data, the better you will be able to use them wisely during the modeling phase.

This aims to precisely determine where we should look for the data and which data we should analyze, and identify the quality of the data available as well as link the data and their meaning from a business perspective.

That means understanding first what we can do with these data before exploring them. This includes some basic knowledge about the methods that will be used and complexities involved.

It seems obvious because data are the main raw material of an efficient data analysis process. So, if you do not understand the nature of the data related to the problem you are trying to solve, consider that you will not be able to solve it.

Formulating some business questions to develop a method is important, such as: which sources do they use? What data to collect? Why these data? To do what? What answers to expect? How much data should be processed? Should we do analysis in real time or periodically?

Therefore, to understand the context of the target problem, you have to play, in some ways, a role of a detective. This can allow you to discover and understand the different elements related to it and determine the tools you need.

It is therefore essential to have at least a basic notion of statistics and mathematics to determine the right analysis technique according to the nature of each data. That also means a significant part of

identifying the technologies that will be most relevant for managing the volume and flow of data.

5.2.3. *Defining the tasks to be accomplished*

The specific tasks to be accomplished corresponds to the problem we are trying to solve by modeling the situation. We can distinguish a number of cases that often recur in a business environment, such as product recommendations.

I will also mention the identification of frauds in the transactions case, the prediction of the impact of a marketing campaign on the conversion rate, or the prediction of the optimal price of a product to maximize the number of sales.

Each task will translate differently and will of course require the choice of different techniques and algorithms.

5.2.4. *Which technology to adopt?*

Merely collecting or having access to large datasets is not sufficient to produce a result. Most of us are not sufficiently prepared for the knowledge extraction process and rapid decision-making. More or less thorough knowledge of at least one analytical tool is usually required.

For data analytics, preference is given mainly to computer languages, which are standardized for data analysis and information extraction. To meet these information-sharing developments, we need tools across the board to help. We need infrastructure and technologies that accommodate ultrafast data capture and processing.

Devices, networks, central processing and software are used to help us discover and harness new opportunities. The key elements to identify are as follows, taking into account the growing advanced analytics tools: which technology to adopt? Why should enterprises adopt this technology?

If some companies do not require it, mastering the Hadoop platform is most often required. Similarly, an experience with Hive and Pig processing tools is an additional argument for recruiting. Cloud tools like Amazon S3 are also important.

According to the 2017 Big Data Analytics Market Study conducted by Dresden Adivsory, big data technologies are adopted by 53% of companies over the world. This report indicates that reporting tools, dashboards, data visualization tools, self-service, data warehouses and real-time data analytics are the most-used technologies in business intelligence (BI).

The processing stages that apply to BI applications also apply to big data, but demand extra technological effort to enable the complete process of data capturing, storage, search, sharing, analytics and visualization to occur smoothly.

Manyika *et al.* suggest that organizational intent should be accorded to the implementation of several new technologies and techniques, in accordance with the big data strategy, in order to extract value from their data.

5.2.5. *Understanding data analytics is good but knowing how to use it is better! (What skills do you need?)*

When working with big data, do not only focus on the technological issues, but also on how we can turn that data into patterns. An algorithm is a "black box"; a user can introduce data (inputs) and they will obtain the results (outputs). How the algorithm works is not the user's business. It is like when someone drives a car without having any idea about its mechanisms, because knowledge is different from know-how.

The key lies in our ability to appreciate the quality and the defects of these algorithms. So, we should rather be interested in questions and issues related to the reliability of the results, their value, etc. It must be considered that there is no perfect model, but models that adapt better to situations are used.

In addition, knowing some analytics methods (decision tree, K-means, etc.) can be a real asset. Since these different techniques can be directly implemented using software (SAS, R, Python, etc.), it is not necessary to know how their algorithms work. The important thing is to understand how they work in general terms and to know which method is most relevant depending on the situation. Selecting the right method for the data that you have is a very important step.

Also, it is essential to know how to manage unstructured data coming from social networks, or video or audio streams. These data are the main challenge of big data.

5.2.6. *What does the data project cost and how will it pay off in time?*

First, you need to determine whether the targeted data give you a return on the investment made by collecting and storing them. Investing in data whose processing cost would be higher than their probable value is indeed to be avoided.

Then, you must also have an idea about some issues such as: how much data is needed for each step? What it is meant to achieve? Any presentation of data analysis results must be accompanied by a summary of the resources committed (investments, etc.) and earnings expectations directly related to the analysis produced by these resources.

For example, in a data analysis process aimed to optimize inventory management, we compare the cost represented by data access, the process time and any external resources, so that this optimization will generate savings for the 12 upcoming months. Questions that are all the more important when the company has invested in software: is there a need for external data? Should a consultant be employed? etc.

5.2.7. *What will it mean to you once you find out?*

Asking interesting questions develops your inherent curiosity about data that you are working on. Knowing something about everything

equips you to understand the context and have the ability to extract the value from data. The key is to think broadly about how to transform data into a form which would help us to find valuable tendencies and interrelationships.

The following types of questions seem particularly interesting:

– What things might you be able to learn from that data?

– How can you ever hope to understand something you cannot see?

– Which techniques and methods do you think will prove more accurate?

– How can you avoid mistakes and get the best models?

– How can you learn lessons by analyzing the available data, and what are you going to do with it?

– How best can you use the results of these analyses?

– What impacts do you expect on the choices to be made?

So, please take note; this involves you, the reader. Seek the correct data to answer a given question. To think like a data scientist or to be a data scientist means thinking about the "meaningfulness" of data, and thus its practice.

5.3. Next steps: do you have an idea about a "secret sauce"?

Big data analysis is a complex process, and we need to focus on extracting useful knowledge from a large amount of different types of data (structured, semistructured and unstructured). This process involves multiple distinct phases that include data acquisition and recording; information extraction and cleaning; data integration, aggregation and representation; query processing; data modeling and analysis and interpretation [SED 17].

Each phase is a separate collection of different techniques and methods in order to improve data analysis results. Some of the phases could be changed, adjusted and repeated, depending on the results, which were provided by other phases.

Valuing the data flows and developing intuitive and innovative ideas requires a solid layer of analysis, because the secret sauce of big data is precisely in this fine diversification of massive data analysis. So, when big data was born, the big analysis of the data, with their different types, is required. Under this massive analysis there are actually hidden tools that exploit and make visible and valuable the mass of data collected.

To be clear, and in order to help you to understand the data analytics process, I would like to remind you that during the Taylorism period, the most effective method for solving any complex problem consists of breaking into simpler subproblems.

We will follow the same context, taking into account that the more of the value of big data is revealed, the more important data analytics process becomes, which can be divided into the following phases.

5.3.1. *First phase: find the data (data collection)*

Data are collected and enriched with the support of advanced technology (sensors, etc.). Moreover, the data are validated in terms of their format and source of origin. Also, they are validated in terms of their integrity, accuracy and consistency.

This phase addresses several analytics challenges (already discussed in Part 1), such as: (1) the complexity of data (collected from different sources and different formats); (2) the challenge of noisy data, as big data generally include different kinds of measurement errors, outliers and missing values; (3) the challenge of dependent data, such as in varied types of current data, such as financial time series and so on; (4) the need to ensure consistency and quality.

Note that data collection presents several peculiarities when compared to traditional data consolidation of distributed data sources, such as the need to deal with heterogeneous flows [SOL 17].

5.3.2. *Second phase: construct the data (data preparation)*

Once the data have been collected, we proceed to the preparation stage. This is not the most enjoyable part of the process, but that does not make it any less essential.

The data preparation phase groups the activities related to the construction of a dataset to be analyzed that is made from the raw data. This includes the classification of data according to selected criteria, the cleaning of data, and especially their recoding to make them compatible with the algorithms that will be used.

It must also be ensured that the data are consistent, without missing values for example. Then, all of these data must be centralized in a database.

Rest assured that you do not need to know the most complex algorithms but you must have a good knowledge about the data and prepare the ground with upstream processing. The important thing is to prepare the ground for the next steps, which will be greatly simplified if this tedious work is done well upstream.

5.3.3. *Third phase: go to exploration and modeling (data analysis)*

The collected, cleaned and prepared data can now be explored. Finally, you can enter the most interesting phase of the analytics process, i.e. the creation of the analytical model associated with the data we are interested in.

This step allows us to better understand the different behaviors and to understand the underlying phenomenon. Feel free to display all kinds of graphs, compare different variables to each other, test correlation hypotheses, clustering, etc.

At the end, you will be able to:

– propose several hypotheses about the causes underlying the generation of the dataset: for example, you will be able to understand if there is really a relationship between two variables X and Y;

– build several possible statistical modeling paths that can help in solving the problem statement;

– introduce, if necessary, new sources of data that would help you to better understand the problem.

This book cannot vouch for your skills, approaches and experiences with managing big data, and any additional important areas that you consider in your product plans. But, as was mentioned at the beginning of the book, readers must have some degree of exposure to probability and statistics. So, I suggest that you take a look at the fundamentals related to these two fields if necessary.

The modeling includes the choice, the parameterization and the testing of different techniques as well as their sequence, which constitutes a model. This process is primarily descriptive to generate knowledge, explaining why things have happened. It then becomes predictive by explaining what will happen, then prescriptive, allowing for the optimization of the future situation.

This phase deals with the structuring, storage and ultimate analysis of data streams. The latter analysis involves the employment of data mining and machine learning techniques such as classification, clustering, etc.

5.3.4. *Fourth phase: evaluate and interpret the results (evaluation and interpretation)*

Before operationalizing the model, you need to evaluate the quality of the model, i.e. its ability to accurately represent our case study, or at least its ability to solve our problem statement. Good results require an effective strategy of data collection, preparation and analysis.

The evaluation verifies the model obtained to ensure that it meets the objectives formulated at the beginning of the process. It also contributes to the decision of the deployment of the model or, if necessary, to its improvement. At this stage, the robustness and accuracy of the models obtained are tested.

It is therefore a game of going back and forth between the modeling phase and the evaluation phase that is carried out to obtain the most satisfactory performance possible. It is even possible in some cases to question certain initial assumptions and to start again in an exploration phase in order to better understand the data.

When the quality of the performance of your model is satisfied, you will be able to move on to the next step, which is the potential deployment of the model in production.

5.3.5. *Fifth phase: transform data into actionable knowledge (deploy the model)*

As part of this phase, the analytics techniques and models identified in the previous phase are actually becoming operational. This phase also ensures the visualization of the data/knowledge according to the needs of the situation.

Hypothetically, suppose that you find that your traffic evaluation model is very powerful and deserves to be shared with more people. You need to deploy it where everyone can get information that will allow them to estimate the traffic according to your model, which will allow them to better orient their journey.

This is the final phase of the process. It consists of a production run for the obtained models' end users. Its aims to mold the knowledge obtained into a suitable form, and integrate it into the decision-making process.

The deployment can also go from the simple production of a report, which describes the obtained knowledge, to the establishment of an application that allows the use of the obtained model for the prediction of unknown values of an element of interest.

The final cycle of the data analytics process sketched above can be schematized in the following way:

Upstream of the analytic approach are data, and downstream refers to knowledge and then action. Therefore, please note that the

importance of the data analytics process includes all of the steps from recovery to deployment.

More of the work is done on recovery, cleaning and data mining, than on modeling itself. It is therefore understood later in this book that these steps have already been performed on the data used.

5.4. Disciplines that support the big data analytics process

The phases outlined previously are supported by a range of data management and analysis disciplines, which are detailed as follows.

5.4.1. *Statistics*

Statistics provides the theory for testing hypotheses about various insights from data. It is intended to match the data with a predefined model whose parameters may vary. The approach generally consists of assuming that the observations follow a known distribution and then testing this hypothesis in order to confirm or refute it.

For example, if you want to model the results of throwing a six-sided die, the procedure will generally consist of making the assumption that the die is balanced (each side has an equal chance of six appearing at each throw), then throw it a number of times, in order to verify this hypothesis. If the model is validated, it means that the probability that this hypothesis is false is sufficiently low given the results obtained. On the other hand, if the probability that the hypothesis is false is too high (results of throws inconsistent with the hypothesis), we will consider another model which we will test in turn.

5.4.2. *Machine learning*

Machine learning (ML) enables the implementation of learning agents based on data mining; ML includes several heuristic techniques. ML is a self-learning method, i.e. an artificial intelligence that allows the machine to produce estimates or forecasts whose

performance will depend on the data. This allows us to say that ML is a discipline at the crossroads of big data and artificial intelligence, which presents a discipline that seeks to solve complex logical problems by "imitating" the human cognitive system.

If we take the example of the die, the approach will consist of launching the die a number of times, then calculating an empirical probability for each result. The higher the number of launches in the learning phase, the better the results.

For more clarity, let us briefly illustrate what machine learning can do with a simple case, probably closer to everyday life: *an anti-spam filter*. At first, we can imagine that the system will analyze how you will classify your incoming mails in spam. Because of this learning period, the system will deduce some criteria of classification.

For example, the probability that the machine will classify a mail in spam will increase if the e-mail contains terms such as "money", "free", "win", etc. and the fact that the sender of the mail does not appear in your address book. On the other hand, the probability of ranking in spam will drop if the sender is already known and the words of the mail are more reliable.

With machine learning, we move on from imperative computing based on hypotheses to probabilistic computing based on real data. So, in addition to the importance of understanding the taxonomy of the system ("IF", "THEN", etc.), we need, first of all, the data.

5.4.3. *Data mining*

Before one attempts to extract and acquire useful knowledge from data, it is important to understand the overall approach or the process that leads to finding new knowledge.

The process defines a sequence of steps (with eventual feedback) that should be followed to discover knowledge in data (see the knowledge discovery process). To advance through each step

successfully, we must apply effective data collection, description, analysis and interpretation [PIE 15]. Each step is usually realized with the help of available software tools. Data mining is a particular step in this process – application of specific algorithms for extracting models from data.

The additional steps in the process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, ensure that useful knowledge is derived from the data.

Data mining and knowledge discovery combines theory and heuristics toward extracting knowledge. To this end, data cleaning, learning and visualization might be also employed.

According to the Gartner Group, this process can be repetitive or interactive depending on the target objectives. We can say that the main task of data mining is using methods to automatically extract useful information from these data and make them available to decision-makers.

5.4.4. *Text mining*

Text mining is the analysis of data contained in natural language text. It refers to the technique that automates the processing of large volumes of text content to extract the key trends and to statistically identify the different topics that arise. The application of text mining techniques to solve business problems is called text analytics. Techniques of text mining are mainly used for data already available in digital format. Online text mining can be used to analyze the content of incoming emails or comments made on forums and social media [SED 16].

5.4.5. *Database management systems*

These systems include relational database management systems, NoSQL databases and big data databases, such as the Hadoop

distributed file system, which provide the means for data persistence and management.

5.4.6. *Data streams management systems*

These systems handle transient streams, including continuous queries, while being able to handle data with very high ingestion rates, including streams featuring unpredictable arrival times and characteristics.

Now that you have understood the context of the data analytics process and you have a basic understanding of the different issues, which you need before approaching any process, it is time to get into the practical side of things. In other words, how about seeing how the model is built?

5.5. Wait, it's not so simple: what to avoid when building a model?

The new business context is driving increased responsiveness to rapidly changing trends and a more rational understanding of an increasingly complex world. More reassuring than intuition, data analysis makes it easier, or faster, to make decisions.

Remember, one of the goals of the data analytics process is to find a model that approximates the reality (the phenomenon) that by using this model we will be able to predict.

A model can be seen as a mathematical function, to which we introduce input "data" that characterize the phenomenon that we want to predict, and that at the exit proposes a score or a result (output) for this phenomenon.

For example, suppose that we are building a model analyzing the number of *likes* in order to answer the question which addresses the probability that one Internet user clicks *like* on this book? Then, we define a value between "0", which refers to a low probability of clicking on *like*, and "1", which presents a maximum probability of clicking on *like*, and thus identify a "chance of clicking on *like*".

Before building the model, there are three essential elements to take into account in our model:

– *Describe*.

The first essential point before designing a model is the description of the phenomenon that will be modeled by determining the question to be answered. The statistical description gives a global view of trends and dominant patterns that structure, for example, the logic of purchase, contact and satisfaction. This first segmentation makes it possible to build a typology of customers and better target an offer.

– *Predict*.

A model can be used to anticipate future behavior. It can be used, for example, to identify the future *likers* of this book. We can then imagine, among the users or the visitors of the website (which presents the database), those who will buy and/or suggest this book, or not. The volume of data delivered proposes a prediction of behaviors at the individual level. Big data offers unprecedented opportunities for segmentation, targeting and identifying new prospects.

– *Decide*.

Prediction tools and models provide insights that can be useful in the decision-making process. Their activities and actions will lead to better results for the company because of the "intelligence" of the model creation process. The model provides answers to questions about the anticipation of future behaviors, or the discovery of a hitherto unknown characteristic concerning a phenomenon, by detection of certain profiles or look-alikes. Recommendation engines prescribe scenarios of possible actions that take into account both predictions and instructions from the client.

In addition, we identify that the data sources are the prerequisite for designing a model. The collected data can be:

– e-mail: open, click etc.;

– web data: browsing, number of likes, etc.;

– offline: store purchases history.

Generally, the more different sources of data are used, the more opportunities generated by the model will be able to detect an interesting element to better analyze the phenomenon.

Multiple methods are available to build a model. The use of statistical models allows for a very good understanding of the relationships between different variables. The choice of the method must therefore be made according to the desired results as well as the completeness of the data. Here are some characteristics to take into account when choosing a model:

– interpretability;

– simplicity;

– accuracy;

– speed (testing and real-time processing);

– scalability.

A good approach is to start with simple models and then increase model complexity as needed, and only when necessary. Generally, simplicity should be preferred unless you can achieve major accuracy gains through model selection.

Once this choice is made and the constraints imposed, the question *how?* is necessary to build the famous model that will get closer to the available data. The models are most represented by a set of parameters that will be put in the vector, which we call β.

For example, a function can be represented by the equation:

$$Y = \beta^T x$$

It is a parametric model and building the model is related to finding the optimal value of β.

It should be mentioned here that there are also non-parametric algorithms, which are not constrained by parameters and suddenly grow in complexity depending on the data.

In the process of building the model, the main notion is that of "loss" of information due to the approximation that we have mentioned above. This determines how modeling the phenomenon, which is an approximation of reality, loses information compared to the reality or the phenomenon.

The more the loss decreases, the closer we get to reality, and the better our model. The vast majority of supervised algorithms (which will be detailed in Part 3 of this book) use this loss function.

There are many ways in which loss can be presented, as described in the sections below.

5.5.1. *Minimize the model error*

A first way of representing this loss is by what is called "error", which refers to the distance between data and the prediction generated by the considered model (see Figure 5.1).

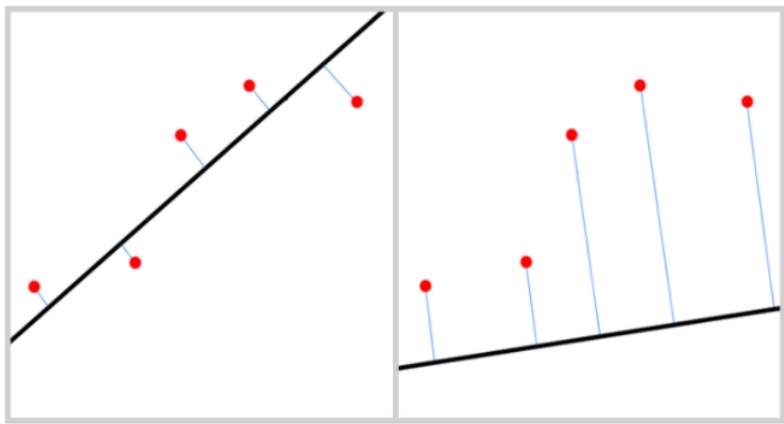


Figure 5.1. *Loss of information*

Regarding the graph on the left side of the figure, we can say that we do not lose too much information. On the other hand, the graph on the right side shows that we are too far from reality

5.5.2. *Maximize the likelihood of the model*

Indeed, the loss in this case is a bit hidden, but we can find mathematically that maximizing likelihood is actually equivalent to minimizing a loss function. The objective is to converge toward the maximum of the likelihood function of the considered phenomenon by finding β from the initial observations.

The loss functions are illustrative examples of the approach that is developed to build a model, because a model is a story of optimization. A large part of the models thus lies in the optimization methods, i.e. the methods that will seek a maximum or a minimum of a determined function.

Once a model is built, we want to use it with new data and new individuals. In practice, the optimization algorithms are built into the model you want to create.

5.5.3. *What about surveys?*

The different models are designed to determine changes or similarities in past behaviors and highlight the most important ideas. Taking the example of the book, we will be able to find, because of a model:

– which users should be targeted first;

– which advantages (reduction, for example) to put in place for a specific customer;

– what type of content (paper version or e-book) and which form of communication (mail, etc.) will allow for the best results.

A model must be used with care, but they are real levers of performance and using these methods allows both a better understanding of the situation within the company and an optimal exploitation of its data.

The algorithm can make sense of the data: identify anomalies, probabilities, anticipate trends, etc. and learn to make increasingly relevant decisions.

Take the example of streaming platforms, when algorithms analyze the content we consume. By comparing our choices to the habits of millions of other users, they determine the type of content that may please us. By integrating the rate of transformation of its suggestions, the machine continually learns from its mistakes and modifies its "reasoning" and becomes more and more relevant.

We suggest other products to buy on Amazon, movies to watch on Netflix, music on Spotify, etc.; this is really useful for both the user and the company who can offer the most relevant content.

Today, new analytics prediction tools can deploy relevant models and enable analytics in applications where insight is needed.

But the multiplication of models poses certain challenges. For example, public opinion surveys use statistical models to draw up sampling plans that can then be used to obtain the best forecasts from respondents' answers.

Many online surveys do not use any method to ensure that respondents are truly representative of the population of interest. It is therefore easy to realize that their result will be biased.

The purpose of a survey of statistical nature is to know the proportion of the population studied that will choose to perform a predefined action. To do this, and because it is often impossible to survey an entire population, we interrogate a small portion of it. The aim in this case is to extrapolate to the scale of the entire population results that have been obtained only from the relatively small number of individuals actually interviewed.

This process is well known in opinion surveys that precede elections. However, the models used in these surveys sometimes produce results that are far removed from reality. This was the case with "Clinton vs. Trump", where most U.S. forecasts nominated Hillary Clinton to be a winner in front of Donald Trump, when it was ultimately the opposite.

The peak of this type of presentation was reached with a forecast from the Huffington Post giving, on November 7, 2016, a more than 98% chance of victory for the Democratic candidate.

Based on a series of specialized software from the NGP Van Company, but also on the files of voters inherited from Barack Obama's campaigns, the technology was to allow the Democrats extremely precise targeting of voters to convince, county by county, neighborhood by neighborhood, and even house by house, with personalized arguments based on the information collected. This software also helps to manage volunteers, identify the most active, position them in the right places, etc.

But these advanced and specialized tools failed to identify the campaign, and misguided it. Several states, including those of the "Rust Belt" (industrial states, such as Michigan and Wisconsin), which were considered to be acquired in advance, and therefore were little "plowed" by the Democrats, finally voted for Trump.

An opinion poll is a survey of public opinion from a particular sample. Opinion polls are usually designed to represent the opinions of a population by conducting a series of questions and then extrapolating generalities in ratio or within confidence intervals.

Generally, the most common method used for opinion polls is quota surveys, or "stratification". The pollster makes an assumption on the total number of individuals belonging to each class of the population, then probes a certain number of these people before generalizing the result to the whole class and applying to it a weighting proportional to the total number of individuals.

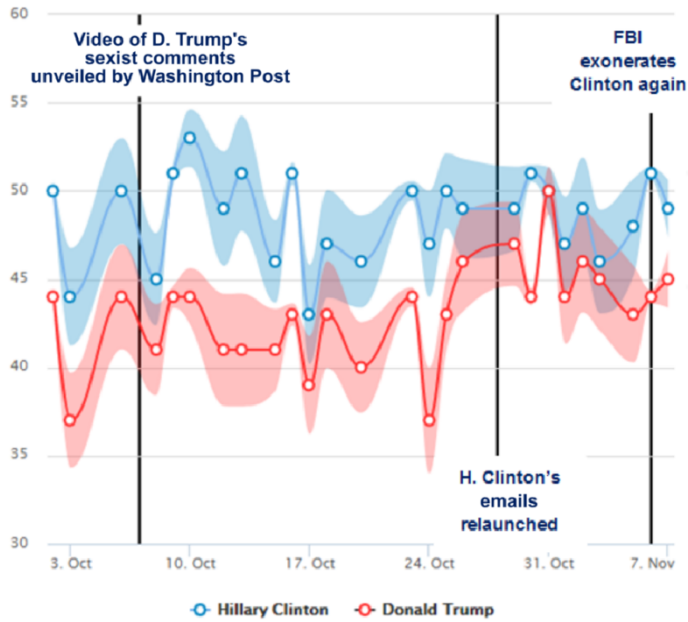


Figure 5.2. *Evolution of the US presidential polls (source: HUFFINGTON POST)*

Based on this hypothesis, the *Los Angeles Times* considered, against all evidence, that all African–Americans between 18 and 21 years, would vote for Trump.

The main cause of these discrepancies lies in the statistical model used, and therefore in the choice of individuals interviewed and the weighting associated with them.

5.6. Conclusions

In conclusion, there is a huge amount of data in today's big data area because most of our activities leave a digital footprint. The analysis of these data will help us to better understand our behaviors (strong points, weak points and improvement points) to better intervene in future because of the analysis models that can be developed in the form of algorithms. For this, particular attention must be paid to the developed algorithms of ML in the third and final part of this book.

PART 3

Data Analytics and Machine Learning: the Relevance of Algorithms

6

Machine Learning: a Method of Data Analysis that Automates Analytical Model Building

> "The greatest value of a picture is when it forces us to notice what we never expected to see".
>
> John Tukey, American Mathematician

6.1. Introduction

With today's objects becoming more and more connected, data collection is global and massive. What is called big data aims to know and evolve our lifestyles, our uses, as well as the way we consume. The ability to process and analyze data is therefore a major issue, and tools having the feature of machine learning (ML) will be particularly useful. That leads to a coordination between "big data" and "ML".

At the crossroads of applied mathematics and computer science, ML consists of the algorithms application to produce prediction and decision support tools based on data. The concept of ML emerges from this environment. Computers can analyze digital data to find patterns and laws in ways that is too complex for human beings.

Data Analytics and Big Data, First Edition. Soraya Sedkaoui.

To explore this area, this chapter discusses the ML context and process as an important aspect of the AI and as one of the main tools for a data scientist. So, welcome to the ML arena.

6.2. From simple descriptive analysis to predictive and prescriptive analyses: what are the different steps?

The application of analytics can be divided, as already detailed in this part two of book, into three main categories, namely descriptive, predictive and prescriptive analytics.

Descriptive analytics involves using advanced techniques to locate relevant data and identify remarkable patterns in order to better describe and understand what is going on with the subjects in the dataset.

Data mining, the computational process of discovering patterns in large datasets involving methods at the intersection of artificial intelligence (AI), ML, statistics and database systems, is accommodated in this category [SUM 06].

Descriptive models can give a clear explanation as to, how and why a certain event occurred, but all of this is already perfectly in the past. So, based on the past, companies can have a clear vision of the future, on what is more important and how they can function. This appeals to predictive models that are seen as a subset of data science [WAL 13, HAZ 14].

Liu and Yang [LIU 17] formalize the way in which a predictive model is made self-organizing via big data. It makes use of available data (several types, created in real time, etc.), statistical methods and various algorithms of ML in order to identify the likelihood of future insights based on the past. The built model predicts by answering the question: what is likely to happen?

Predictive analytics use data, statistical algorithms and ML to predict the likelihood of business trends and financial performance based on their past behaviors. They bring together several

technologies and disciplines such as statistical analysis, data mining, predictive modeling and ML technology to predict the future of businesses.

With the increasing number of data, computing power and the development of AI software and simple analytical tools' uses, many companies can now use predictive analytics. For example, it is possible to anticipate the consequences of a decision or the reactions of customers.

Predictive analytics is the act of predicting future events and behaviors present in previously unseen data using a model built from similar past data [NYC 07, SHM 11]. It has a wide range of applications in different fields, such as finance, education, healthcare and law [SAS 17].

In this case, it should be mentioned that the amount of data available is not the problem; the richness of the data however is often questionable. This is most certainly required when people want to perform prescriptive analytics.

When executed correctly, this application of mathematical and computational algorithms enables decision-makers to look into the future of their own processes and opportunities. It even presents the best course of action to take for gaining advantages.

The requirements for an accurate and reliable prescriptive analytics outcome are hybrid data, integrated predictions and prescriptions, taking into account side effects, adaptive ML algorithms and a clear feedback mechanism.

AI and ML can be considered as a top level of data analysis. Cognitive computer systems constantly learn about the business and intelligently predict industry trends, consumer needs, etc. The level of cognitive applications can be defined by four main skills:

1) an understanding of unstructured data;

2) the ability to extract information and ideas;

3) the ability to refine expertise with each interaction;

4) the ability to see, hear and speak in order to interact with humans in a natural way.

Along with mathematical, statistical and analysis methodologies, ML and big data analytics have emerged to build systems that aim at automatically extracting information from the raw data that the IT infrastructures offer.

6.3. Artificial intelligence: algorithms and techniques

Artificial intelligence found its name at the Dartmouth conference in 1956. But it began in the early 1950s with, for example, the work of Alan Turing which questioned whether we could make a computer think. He proposed a test called "the Turing test", in which a person chats through a computer and must guess if their interlocutor is a machine or a human being.

If the person cannot pass the test, then we can conclude that it is possible to operate a computer with logic algorithms similar to our way of thinking or even beyond. Indeed, AI has vegetated several times, especially in the 1970s and 1990s; because it has been limited for a long time by the costs and performances of the machines (speed, memory capacity, storage capacity), which have undermined the expectations that had been placed there, causing frustrations and losses of investments by industrialists.

AI had its beginnings in computer science with automated systems, recurrence and languages like "Lisp" and "Prolog". In its early days, we mainly talked about logical rules, recursion, parsing, graphs and expert systems.

Today, IT advent has abolished these limits and we are witnessing an explosion of its possibilities that seem unstoppable. The clusters of OVH, Amazon AWS and Google Cloud servers available with an individual's budget are good examples.

Most of the techniques used in AI are based on mathematical theories (advanced statistics, decision trees, Bayesian networks, neural networks, etc.) that have been known for 50 years or more. Now, the techniques used to make our machines think are many:

– fuzzy logic;

– genetic algorithms;

– data mining;

– Bayesian inference;

– smart agents;

– neural networks;

– automatic learning.

They are becoming more and more used today because of the conjunction of a number of factors:

– data storage costs are constantly decreasing;

– the increase of computing power;

– the explosion of the amount of information available in digital form;

– this information is largely unstructured and requires operating techniques different from conventional methods.

So, we can run complex calculations and analyze billions of data for derisory costs. This boosts research and gives rise to AI. This opens the way to ML, which takes all its dimensions with very large volumes of data. So, the arrival of big data has propelled a new field of AI: ML. And this one already gives fabulous results that boost all the investments of sectors: research centers, banks, insurance, finance, aerospace, automobile, pharmacy, etc. The whole industry seizes it.

ML is a data analysis technique that teaches computers what humans are naturally capable of learning from their experiences. ML's algorithms use computational methods that "learn" information

directly from the data without the need to rely on a predetermined equation as a model. Algorithms adapt and become more efficient as the number of samples available for learning increases.

Algorithms of ML identify natural patterns in the data that generate useful information and help to make better decisions and predictions. They are used daily to make critical decisions in medical diagnostics, stock trading, energy load forecasts and more. For example, websites take advantage of ML to process millions of options to recommend songs to listen to or movies to watch. Retailers use it to understand the buying behaviors of consumers.

With the rise of big data, ML has emerged as one of the best problem-solving techniques in some areas, including the following:

– *finance*: banks and insurances use ML to discover important information within the data and to prevent fraud. Also, for credit evaluation and algorithmic trading;

– *health*: ML is being used more and more in the healthcare industry, particularly with the rise of connected objects and other sensors that make it possible to use the data to access a patient's health data in real time. This technology can also help medical experts analyze data to identify alarming trends to improve diagnostics and treatments;

– *marketing*: websites that recommend products based on a user's previous purchases use ML to analyze the customer's purchase history and offer products that might interest them. The ability to collect, analyze and use data to personalize the shopping experience represents the future of retail;

– *image processing and computer vision*: they are used for facial recognition, motion detection and object detection, or recognition of friends in photo albums or via search engines. Also, voice recognition (during a phone call) to dictate a text to your smartphone or to recognize a song on the radio (Shazam, SoundHound applications, etc.). There is also the case of Apple with face recognition in the iPhone X;

– *biology*: ML is used for tumor detection, drug discovery and DNA sequencing;

– *energy production*: it is used for forecasting prices and charges. ML can also find new sources of energy, analyze minerals in the soil, or predict sensor failures in refineries. This technology makes oil distribution more efficient and economical;

– *automotive, aerospatial and industrial production*: they are used for predictive maintenance or for the automatic control of our cars, etc.;

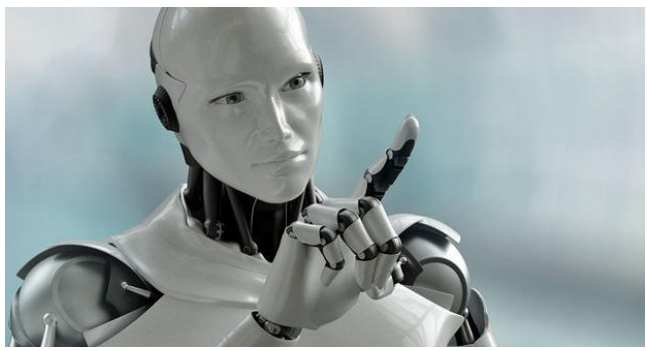
– *natural language processing*: it is used for speech recognition applications;

– *art* also does not escape: it is used for making music, poetry or even paintings.

Extremely complicated and expensive disciplines, such as voice recognition and visual/facial recognition, formerly reserved for industrialists and the army, have become part of our daily lives.

6.4. ML: what is it?

When we hear about ML, or more generally about AI whose ML is a subdomain, you may generally think of robots:



Wait... it is not really that, moreover even the robot in the picture confirms it.

Born from pattern recognition, ML refers to all the approaches that give computers the ability to learn autonomously. These approaches,

which overcome strictly static programs for their ability to predict and make decisions based on the data input, were used for the first time in 1952 by Arthur Samuel, one of the pioneers of AI, for a game of checkers. Samul defines ML as the field of study aimed at giving a machine the ability to learn without being explicitly programmed.

Tom Mitchell of Carnegie Mellon University proposed a more precise definition:

> "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E".

Then, he has illustrated this definition with some examples that have been grouped in Table 6.1.

Application	Checkers learning	Handwriting recognition	Robot driving learning
Task "T"	Playing checkers	Recognizing and classifying handwritten words within images	Driving on public four-lane highways using vision sensors
Performance measure "P"	Percent of games won against opponents	Percentage of words correctly classified	Average distance traveled before an error (as judged by human overseer)
Training experience "E"	Playing practice games against itself	A databases of handwritten words with given classifications	A sequence of images and steering commands records while observing human driver

Table 6.1. *ML examples as illustrated by Mitchell*

The basic idea of ML is that a computer can automatically learn from experience [MIT 97]. Using collected data, a ML algorithm finds the relations between different properties of the data. The resulting model is able to predict one of the properties of future data based on properties [ECK 07].

Although ML applications vary, its general function is similar throughout its applications. The computer analyzes a large amount of data, and finds patterns hidden in it. These patterns are mathematical in nature, and they can easily be defined and processed by a machine.

Although it is currently boosted by new technologies and new uses, ML algorithms have been widely adopted in different fields such as business, computer science and so on. To learn and grow, computers need data to analyze and train. In fact, big data is the essence of ML, and ML is the technology that makes full use of the big data potential.

For the analysis of such amounts of data, ML is much more efficient than traditional methods in terms of accuracy and speed. For example, based on data associated with a transaction, ML can detect potential fraud in a millisecond. Thus, this method is much more efficient than traditional methods for analyzing transactional data or data from social networks or CRM platforms.

Their extensive application is due to their ability to automatically extract information from the data. In the context of IT, ML techniques have been used to solve many problems related to anomaly detection, patterns discovery, profiling, etc.

For example, by analyzing website content, search engines can define which words and phrases are the most important in defining a certain web page, and they can use this information to return the most relevant results for a given phrase search [WIT 16].

Another great example of ML applications is the IBM Watson which saved the life of a woman dying from cancer [BOR 16]. The Watson computer system ran her genomic sequence and found that she had two strains of leukemia instead of simply the one discovered by doctors. This enabled a substantiated cure.

We can cite as well, the case of the *Google DeepMind* algorithm that recently learned to play 49 *Atari* video games by itself. In the past, development was limited by the lack of available datasets, and by its inability to analyze massive amounts of data by seconds.

The more an ML system receives data, the more it learns and the more accurate it becomes. Today, data are accessible in real time. This allows AI and ML to move to a data-driven approach. The machine is now sufficiently agile to access huge datasets and analyze it. In fact, companies are now joining Google and Amazon to implement AI solutions for their businesses.

AI tools (mainly natural language processing, pattern recognition and ML) are already very present, even though many people are not aware of this. They are found in everything that can identify voices and faces or answer questions like Apple's *Siri* or Amazon's *Echo*.

MetLife, one of the world's insurance leaders, uses this technology and big data to optimize its business. Speech recognition has allowed MetLife to improve the tracking of accidents and to better measure their consequences. Claims processing is now better supported because claim templates have been enriched with unstructured data that can be analyzed by ML.

Also, this technology is able to learn the habits of the home's occupants. Designers of connected objects, including thermostats, can analyze the temperature of the dwelling in order to understand the presence and absence of occupants so that it may turn off the heater and turn it back on a few minutes before they return.

Then, big data accelerates the learning curve and automates data analysis. ML is ideal for exploiting the hidden opportunities of big data. This technology makes it possible to extract value from massive and varied data sources without having to rely on a human.

It is the science of getting computers to act without being explicitly programmed. ML is driven by data, and fits the complexity of the huge data sources. Unlike traditional analytical tools, ML can also be applied to growing datasets. The more data injected into an ML system, the more the system can learn and apply the results to higher quality insights. ML is thus able to discover patterns buried in data more effectively than humans.

ML improves diagnostics, predicts better outcomes and is revolutionizing personalized care. The basic idea of any ML process is to train the model, based on some algorithm, to perform a certain task: classification, clustering regression, etc.

ML can be used to reveal a hidden class structure in unstructured data, or it can be used to find dependencies in a structured data to make predictions. So, as part of the job of the ML process, it is supposed that a relationship exist between available data and it is the role of algorithms to reveal it. To reveal this relationship, which can lead you to extract value, you must adopt the algorithm that will allow you to do this.

6.5. Why is it important?

The objective of ML is to learn from data, in another word, to learn from real observations. As we have previously noticed, these data can come from different sources and in different natures and forms. Depending on the case, they are more or less complex to analyze. In this order, algorithms aim to extract some regularity that will allow learning. To use the algorithms, the data must be formatted under a matrix representation.

In the ML context, every observation is described by a set of variables: X_j ; $(j = 1,2, \dots n)$. Obviously, all the interest of ML algorithms will be to find regularities in the data through the observation of a large number of individuals (from 1 to m). The value of the variable X_1 of the individual 1 is noted: X_{11}. The general case is thus: X_{ij}, that is to say the value of the variable X_j of the individual X_i.

These n variables describing m individuals are represented in what is called a matrix:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}$$

We can organize them in the form of a table, considering that each column corresponds to a variable and each line corresponds to an individual (Table 6.2).

Individuals	Variables			
		x_1	..	x_n
	1	x_{11}		x_{1n}
	⋮	⋮		⋮
	m	x_{m1}	..	x_{mn}

Table 6.2. *Matrix in a table form*

Note that if the table only has one column, then this means that we only have one variable, meaning we are talking about vector and not matrix.

Whatever the algorithm used, supervised or unsupervised algorithms (we will talk about this again in detail when presenting the ML algorithms in Chapter 7), the goal will always be the same.

Peter Sondergaard, Executive Vice President at Gartner, explains that: "to help companies to develop the digital aspect of their business, the key is to leverage the algorithms". So, value does not lie in big data, of course data is needed, but it is ephemeral, and in itself it will not be transformative. "Data are inherently passive, it's useless unless you know how to use it, how to act on it, because the real value lies in the algorithms, the algorithms define the action", he confirms.

The profitability of big data lies largely in the ability of the company (how?) to analyze the amount of data in order to generate useful information. The answer is "machine learning".

The most obvious answer is that the relevance and performance of the algorithms set up the real value of an analysis system integrating ML capabilities. In other words, those who discover logical principles in a sea of disparate data can monetize these principles.

Basically, there are some enterprises for which algorithms have provided a real way to stand out from the competition, and some other enterprise for which ML algorithms present a practical tool to rationalize their costs. The value of ML algorithms will therefore depend on the objectives set.

Everyone intuitively understands the role of algorithms that improve Google's ability to refine results, target content and monetize ads. It is also easy to imagine that algorithms can help e-commerce sites make purchase recommendations and adjust prices to maximize profits.

Also, it will not be missed by some that video game publishers analyze the behavior and performance of players and rely on ML to encourage in-app purchases.

Then, the big data/ML duo is what enables companies to create added value through their available data. We might ask why. To better understand it, let us already establish that ML is at the heart of the tools we use every day.

Several examples of this duo can be mentioned, such as: the detection of spams from our mailboxes, Amazon's recommendations in the e-commerce area, IBM Watson, taking notes in the medical field, etc. In March 2016, the AlphaGo 4-1 victory (a software developed by Google) against Korean Go champion Lee Sedol is an example of the extraordinary capabilities of ML.

ML goes even further than data mining. The latter sets trends to understand and anticipate and then humans make decisions. With ML, decision-making becomes artificial: there is no human to make decisions; it is the machine that decides automatically.

In fact, ML is a sort of system programmed by humans that allows the machine to operate autonomously. Its strength is that it relies on algorithms to process data and to learn rules as and when there is an opportunity to. So, this AI must have human supervision to accomplish business objectives.

6.6. How does ML work?

In the AI field, which aims to make machines perform tasks that normally require human intelligence, ML is currently the dominant trend. With ML, the computer performs tasks for which it has not been explicitly programmed to complete by producing models itself and sometimes even changing them from new data.

ML algorithms use large amounts of data. They are getting closer to data mining or BI (business intelligence). However, data mining is limited to making data intelligible by presenting them analytically and synthetically. ML goes further by producing rules or models that can explain the data, thus potentially predicting new data (predictive analytics), or even ultimately making decisions based on new data and the established model.

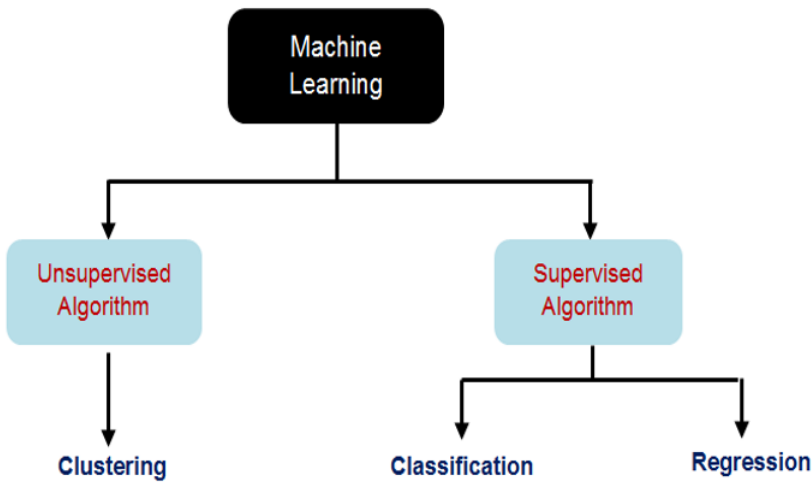


Figure 6.1. *Supervised and unsupervised ML algorithms*

It should be noted, before speaking about the ML process, that ML relies on two types of techniques: supervised learning, which involves training a model on known input and output data, so that it can predict future outcomes, and unsupervised learning, which identifies hidden models or intrinsic structures in the input data.

Working with ML algorithms means that a whole workflow is taking place. It will include the following.

6.6.1. *Definition the business need (problem statement) and its formalization*

Identify the main learning problem by considering what is observed and the answer you want the model to predict. Defining the problem to solve helps to clarify ideas. So, the first step is to imagine a path between the initial data and the value to be predicted.

At this stage, we can describe the problem in an *informal* way. In other words, we mean to formulate the problem in a precise and concise sentence: for example, I am looking for a solution capable of estimating the prices of cars.

Moreover, here we can, for example, translate the definition given by Mitchell to identify: T, P and E. In this case, we can say, for example:

– *task (T)*: the estimate of the prices of cars;

– *performance (P)*: the precision of the algorithm prediction and how close it is to the real price of cars;

– *experience (E)*: the description of the car and its actual price.

6.6.2. *Collection and preparation of the useful data that will be used to meet this need*

Any ML process is largely based on data. This step consists of collecting and registering the type of data useful for solving the problem. Collect, clean and prepare data so that they can be consumed by ML model training algorithms.

You have to know here if your data are ready for analysis. Before the algorithm can provide you any answers, you need to give it raw materials to work with. You need to prepare your data to ensure that it meets the basic criteria for analysis processing. So, you need relevant, connected, accurate and sufficient data.

This is to define what data need. And know the sources from which to draw. It can be external APIs (Facebook, Twitter, etc.), databases, and data files (Excel, XML, etc.), etc.

This is not the only parameter to take into account as it is important to question the quality of these data (are they accurate, are they biased according to particular conditions? etc.).

6.6.3. *Test the performance of the obtained model*

Several types of problems are solved by the ML algorithms, but it is necessary to choose the algorithm that allows us to better model the problem, including classification, regression and clustering. With each type of problem, one can have several candidate algorithms to solve it. The factors that can come into play in choosing the right algorithm can be many, including the number of features, the amount of data we have, etc.

After rolling out its algorithm on its training set and making predictions with the test set, it is time to evaluate its performance. There are some standard ways to do this depending on the types of problems. For example, for regression a model can be considered good by an indicator: R^2.

6.6.4. *Optimization and production start*

The results of an ML problem are often complex to interpret. Whatever the method adopted in the analysis process, the following points should be answered:

– *the context*: raising the context of the problem and the reasons for its resolution;

– *the problem*: concisely describe the problem we are trying to solve;

– *the solution*: describe the solution provided in terms of architecture, how to exploit the solution, etc.;

– *limitations*: if the solution is not universal or has limitations, it is better to list them. This gives a solution credibility and can open paths to new areas of improvement;

– *conclusion*: quickly revisit the description of the problem as well as the solution and the benefits derived from it.

The spam classifier is one of the first ML applications to solve a real-life business problem, and it is incorporated into most of today's e-mail applications [GUT 15].

Another very important axiom to remember when starting up a new ML process is offered by American mathematician John Tukey, who is often revered in statistics circles for his many contributions to statistical methods as well as his book on *Exploratory Data Analysis* published in 1977:

> "The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data".

This implies that ML practitioners need to know when to give up, when the acquired data are not sufficient to answer the question you are trying to analyze.

Also, to harness the power of ML to use data to make better decisions, several automatic solutions have been developed to simplify the process. Python, Matlab, R, and other tools and big data management features offer applications that make ML accessible, making it the perfect environment to apply ML to the data analysis.

Wait just a minute… did you just get a really strong sense of *déjà vu*?

This might be because the process is similar to the big data analytics process described in Part 2 of this book.

Maybe you will now better understand why ML is currently widely used for big data. We have already talked about the duo of big data/ML. It is necessary to combine several ingredients that must be cleverly mixed to succeed with this duo:

– data quality;

– computing power;

– algorithms;

– and talent.

6.7. Data scientist: the new alchemist

Big data analytics is not a modern trend, because data has existed and has been used over time and it is constantly growing. However, this progression is significantly more obvious with the data revolution movement. Data analysis, when it is not preceded by the word "big", refers to the development and sharing of useful and effective models.

Big data as well as traditional analytics search for extracting value from datasets. The added value of big data is the ability to identify useful data and turn it into usable information by identifying patterns, exploiting new algorithms, tools and new project solutions. So, the move toward the introduction of big data and analytics tools within businesses addresses how this new opportunity can be operationalized.

I especially do not want to disappoint you, but *machines* cannot solve all our problems, even if they are capable of solving certain things. For example, no algorithm can start your car. Nevertheless, if we have enough data on the operation of our vehicle, an algorithm can detect a failure and suggest us to change the starter. Maybe it can even suggest it before the starter goes dead!

That is why it is necessary to translate every problem into elements that can be processed and analyzed by algorithms. This can help us to obtain better results and make better use of available data.

When data analysts or data scientists encounter a set of data, they need to understand not only the limits of the data, but also the limits of the questions that it can respond to, as well as the range of possible appropriate interpretations. This includes mathematical and statistical know-how, data modeling, data mining, soft skills, programming and general business acumen.

Another point to take into account is that, when working with big data in a business context, is to consider an investment in security. So, what is really needed is a "data scientist" (which the *Harvard Business Review* has called "the sexy new job of the 21st century") who can understand analytics and has a strong creative streak in order to ask the right questions get significant values from data.

The fields of a data scientist draw on many disciplines. The one who combines analytics and soft skills, with an awareness of business needs and marketing fundamentals, and an appetite to make data visible and understandable, is the one called a data scientist. This versatility is often represented by a Venn diagram (Figure 6.2), which remains a good compass to be locate us in this galaxy of disciplines.

Being a data scientist means being at the heart of data valuing and intervening at all stages of the data chain: problem definition, data collection, preparation, modeling and algorithm creation. A data scientist must know how to present and prioritize the results to be used by decision-makers. Excellent communication skills are also needed.

This scientist is not considered to be an analyst, developer or statistician, but all at once. The data scientist must seek information where it is not. The most popular way to do this is probably to ask a lot of questions in order to see what sticks. But they cannot be just any questions; they have to be significant questions.

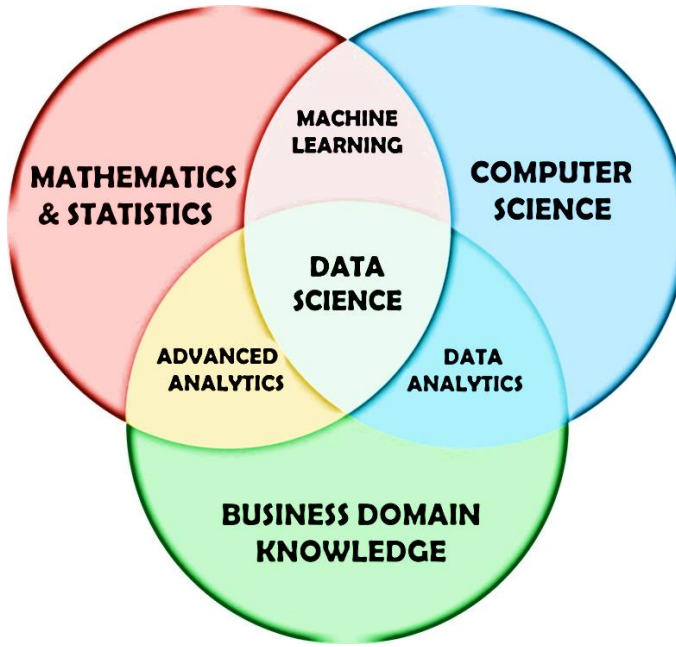


Figure 6.2. *The skill set of a data scientist*

6.8. Conclusion

Now you know how to identify the raw material for data analysis and ML processing and you have an idea about what you can and want to do with it. Now, you are probably eager to exploit it with the many algorithms designed for it. As such, Chapter 7 details a general idea of what all these algorithms can do.

7

Supervised versus Unsupervised Algorithms: a Guided Tour

> "Once we know something, we find it hard to imagine what it was like not to know it".
>
> Chip & Dan Heath
> Authors of *Made to Stick* and *Switch*

7.1. Introduction

In this chapter, we will detail the necessary concept to develop an effective roadmap for implementing a supervised and unsupervised ML algorithm. Here, we will learn how to transform your business objectives into a data analysis process using the ML process. Also, we will discover different techniques for advanced supervised and unsupervised algorithms, such as clustering, classifications and regression models.

This chapter addresses many methods that have their bases in different fields. In other words, this chapter will lay the foundations in order to allow you to grasp the global view, the famous "big picture", which will help you to choose the best algorithms.

Data Analytics and Big Data, First Edition. Soraya Sedkaoui.

7.2. Supervised and unsupervised learning

The goal of ML is to train algorithms so that they can learn to make predictions on a large amount of data. Depending on the type of input data, machine learning algorithms can be divided into supervised and unsupervised learning. Supervised and unsupervised learning comes from data mining, which aims to extract knowledge from databases.

7.2.1. *Supervised learning: predict, predict and predict!*

If an algorithm is given a set of inputs (features vectors): $\{x_1, x_2, \ldots x_n\}$, and a set of corresponding outputs (labels): $\{y_1, y_2, \ldots y_n\}$, then the goal of the algorithm is to learn to produce the correct output given a new input; this is a supervised learning task.

In supervised learning, input data comes with a known class structure [MOH 12, MIT 97]. The algorithm is usually tasked with creating a model that can predict one of the properties by using other properties. After a model is created, it is used to process data that has the same class structure as input data.

A *supervised learning* task is called "classification" if the outputs are discrete or "regression" if the outputs are continuous.

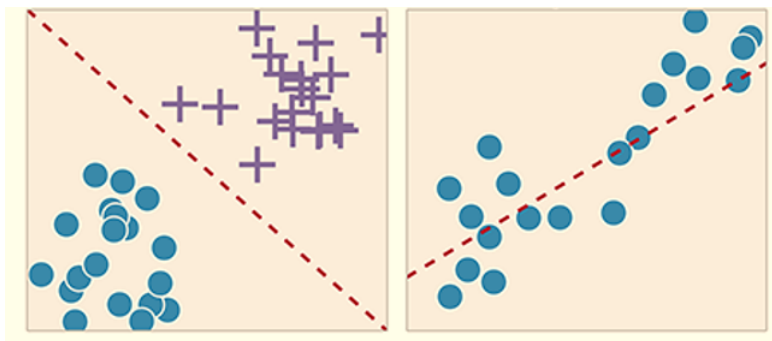


Figure 7.1. *Illustration of classification and regression*

Inputs can be vectors of different types of objects, integer numbers, real numbers, strings or more complex objects. Outputs take values, each representing a unique state.

For example, an algorithm may be given a number of vectors representing numerically external features of a person, such as sex, age, income, etc., and corresponding outputs that take one value from the set “male, female”.

Supervised algorithm can also be applied in the detection of spam from your mail as well as in the forecast scores and risks associated with insurance.

But, how does it work? Let us take a simple example:

We will propose a series of pictures, knowing that the target that we want will be the “category”.

Each picture used to drive the algorithm is grouped in its category:

The objective is then to classify, by means of classification methods, the group of membership of each picture according to its similarity to other pictures.

In supervised learning, you will retrieve annotated data from their outputs to train the model, i.e. you have already associated them with a label or a target class and you want that the algorithm be able to predict it for new non-annotated data once trained. So, the system learns to classify according to a predetermined classification model and known examples.

Supervised learning is divided into two parts:

1) the first is to determine a tagged data model;

2) the second consists of predicting the label of a new datum, knowing the previously learned model.

Supervised learning will be applied when the goal is to predict a value or belonging.

7.2.2. *Unsupervised learning: go to profiles search!*

If an algorithm is only given a set of inputs: $\{x_1, x_2, \ldots x_n\}$, and no outputs, this is an *unsupervised learning* task. Unlike supervised learning, which attempts to find a model from labeled data: $(X) \rightarrow Y$, unsupervised learning takes only untagged data (no variable to predict Y). In unsupervised learning, input data do not have a known class structure, and the task of the algorithm is to reveal a structure in the data [SUG 15, MIT 97].

An unsupervised learning algorithm will find patterns or structuring in the data. In unsupervised learning, there is no initial labeling of data. Here, the goal is to find some pattern in the set of unsorted data, instead of predicting some value. Unsupervised methods usually generate too many false alerts, so it is often a good idea to combine both supervised and unsupervised methods, as in [KRI 10].

Unsupervised learning can be thought of as finding patterns in the data and beyond what would be considered as pure unstructured noise. In unsupervised learning, input data are not annotated. So, *how can this work?* Good question.

The algorithm must discover by itself the pattern according to the data. The algorithm applies in this case to finding only the similarities and distinctions within these data, and then grouping together those that share common characteristics.

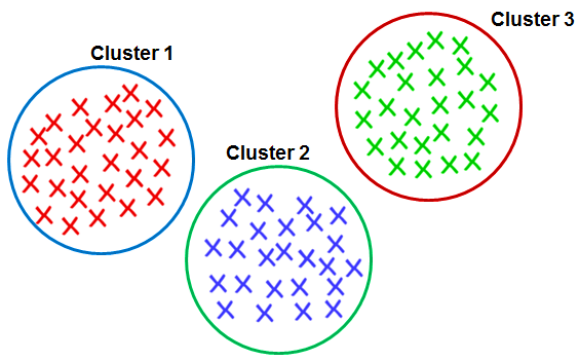


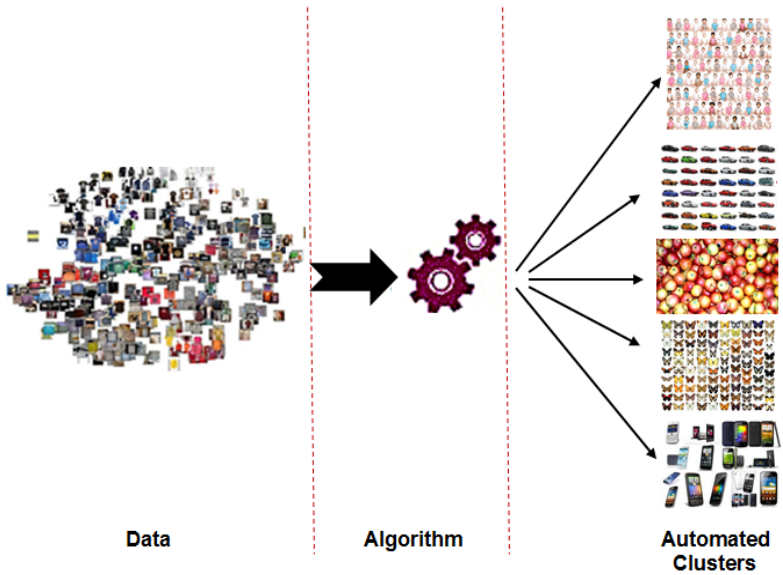
Figure 7.2. *An example of clustering*

Clustering algorithms fall into the unsupervised learning category. They make it possible to group together similar data. Find the hidden patterns in the unlabeled data and separate it into clusters according to similarity.

An example can be the discovery of different customer groups inside the customer base of the online shop.

For example, in the case where Amazon receives a new purchase proposal from you (as a new user), Amazon users are divided into groups and, according to your purchase choice, you will be associated with a group of clients who have purchases close to yours. It is just about bringing clients into groups that are not predefined.

Relating this back to our previous example of the categorized pictures, if we inject thousands more photos, similar pictures would be automatically grouped within the same category.



Also, there could be the case of an epidemiologist, who studies liver cancer victims and wants to try to come up with explanatory hypotheses. The machine could differentiate several groups, and then associate them with various explanatory factors.

Quoc *et al*. [QUO 12], researchers at "Google Brain", applied unsupervised learning algorithms a few years ago on YouTube videos to see what this algorithm could learn. They train a nine-layered locally connected sparse auto encoder on a large dataset of images (the model has one billion connections; the dataset has 10 million (200 × 200 pixel) images downloaded from the Internet).

The results show that it is possible to train a face detector without having to label images as containing a face or not. Control experiments show that this feature detector is robust not only to translation but also to scaling and out-of-plane rotation. They also conclude that the same network is sensitive to other high-level concepts such as cat faces and human bodies.

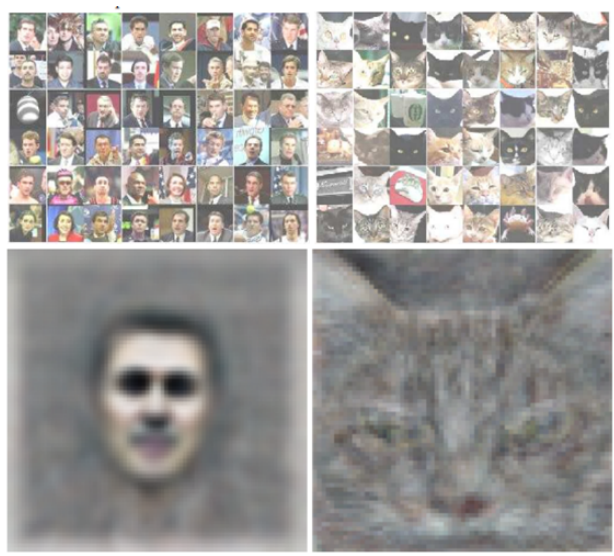


Figure 7.3. *The internal representation of the concepts: "face" and "cat" learned by an unsupervised algorithm (source: [QUO 12])*

7.3. Regression versus classification

Supervised learning assumes the availability of labeled samples, i.e. observations annotated with their output, which can be used to train a learner. In the training set, we can distinguish between input

features and an output variable that is assumed to be dependent on the inputs.

The output, or response variable, defines the class of observations, and the input features are the set of variables that have some influence on the output and are used to predict the value of the response variable. Another distinction that will help you in the choice of a machine learning algorithm is the type of output expected from our program: is it a continuous value (a number) or a discrete value (a category)?

The first case is called a regression, and the second is called a classification. The first assumes a categorical output, while the latter a continuous one. So, depending on the type of output variable we can distinguish between two types of supervised task: (1) classification and (2) regression.

For example, if you want to determine the cost per click of a web advertisement, you apply a regression. If you trying to determine if a photo is of an apple or a banana, you apply a classification analysis.

The regression/classification distinction is about supervised algorithms. It distinguishes two types of output values that can be sought to be processed.

7.3.1. *Regression*

Regression method takes a finite set of relations between dependent variables and independent variables and creates a continuous function to generalize these relations [WAT 16]. Regression predicts the value based on previous observations, i.e. values of the samples from the training set. Usually, we can say that if the output is a real number or is continuous, then it is a regression problem.

For example, let us suppose that you want to predict the income of clients and their prospects based on data such as their socioprofessional category, age, gender, occupation, address and so on. You can collect many observations by conducting a survey on a panel of clients. Then,

some of these observations will be used to generate a model that can predict this income.

The remainder of the panel will be used to measure the accuracy of the algorithm by comparing actual and predicted values. Finally, you can estimate the income and prospects of those clients who are not in the panel.

In the context of a regression problem, Y can take an infinity of values in the continuous set (denoted $Y \in \mathbb{R}$). These might be temperatures, sizes, GDP, unemployment rates, incomes or any other types of variables that do not have finite values *a priori*.

Regression is a simple first example of how an algorithm can learn a model. On the basis of the bank example discussed above, the question we will try to solve is: Given the characteristic of the bank clients' regular monthly income, how much should they normally save?

Imagine that the only characteristic we have is the clients' monthly income. Our training set is n income observations and their associated savings: (x, y) = (income, savings).

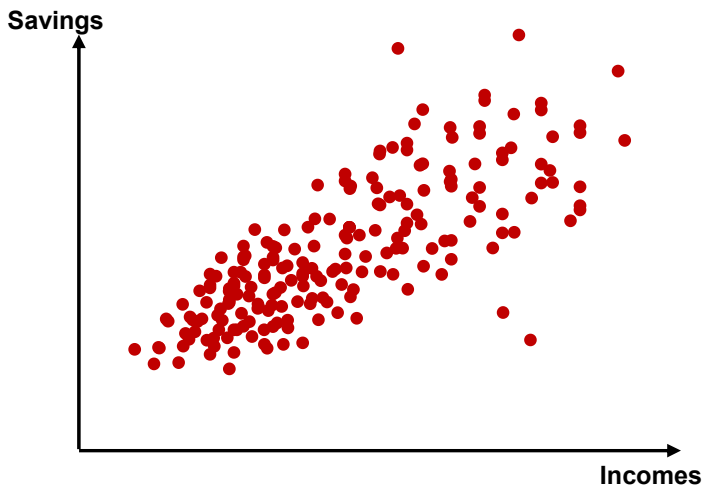


Figure 7.4. *Savings level by income*

Figure 7.4 depicts a two-dimensional graph that shows the relation between income and the dependent variable indicating the clients' saving level.

Clearly, from the visualization of the figure, we can say that the level of savings linearly depends on income. We can therefore extract a modeling hypothesis that the phenomenon has the form of a straight line.

The linear regression is based on the assumption that the data come from a phenomenon that has the shape of a straight line, i.e. there is a linear relationship between the input or the observations and the output, which take the form of predictions.

Then, we have our underlying model constraint which must be in this form:

$$\hat{Y} = \beta^T x$$

with: $x = (1, x_1, x_2, \dots \dots x_n)$ and $\hat{Y} = (\hat{y}_1, \hat{y}_2, \dots \dots, \hat{y}_{n1})$.

The observation vector x starts with 1 because we need the ordinate at the origin. We nominated $\hat{Y}$ in order to distinguish it from real observations. We are talking about the estimate given by the model.

For the simple linear regression, we note: $\hat{Y}_i = \beta_0 + \beta_1 x_{i.}$

The goal is to find the line parameterized by $= (\beta_0, \beta_1)$, which fills the training data better.

We can then graphically represent the regression equation we found in order to verify that it fits well with the data (observations).

Now, we have our parameter β, that is to say we have found the line that best fills our training data; we can make predictions on new data, that is to say predict savings by directly applying a given income as an input into the model.

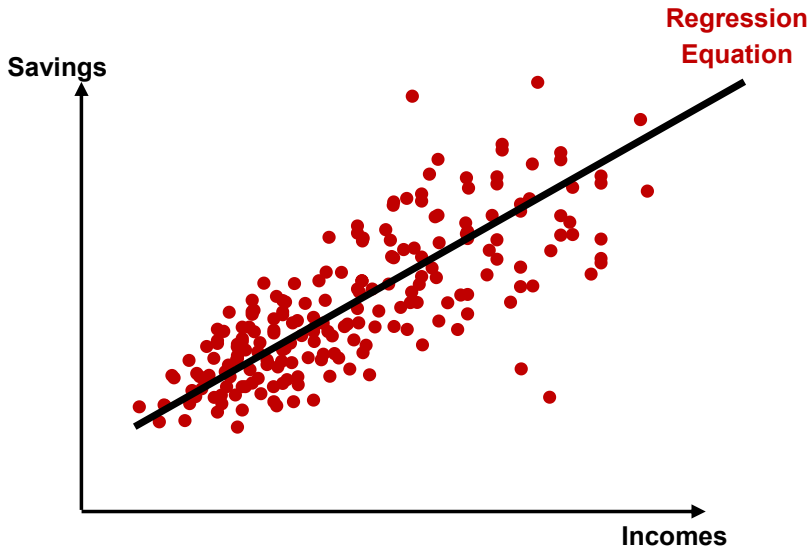


Figure 7.5. *Regression model*

7.3.2. *Classification*

In contrast with regression problems, when the explained variable is a value in a finite set, it is referred to as a supervised classification problem. This amounts to assigning a label to each observation.

Classification is a particular supervised learning task.

In the context of this method, Y takes a finite number k of values: $(Y = \{1, \ldots, k\})$. This is called tags assigned to the input values. This is the case of: "true/false" or "passed/failed". This method can also be used, for example, in health risk analysis. A patient's vital statistics, health history, activity levels and demographics can be cross-referenced to score (the level of a risk) and assess the likelihood of illness.

When the set of possible values of a classification exceeds two elements, we speak of multiclass classification. Figure 7.6 illustrates both types of classification.

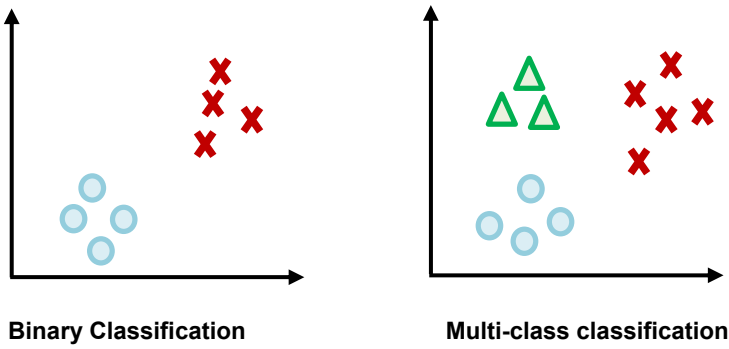


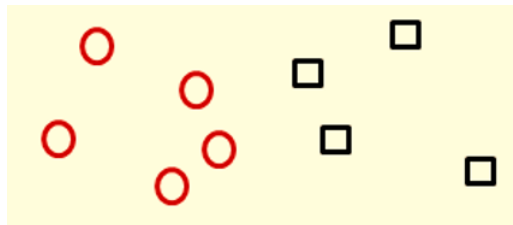
Figure 7.6. *Classification types*

Among the classification algorithms, we find the K-nearest neighbors (kNN), logistic regression, Support Vector Machine (SVM), Naïve Bayes, decision tree, Random Forest and neural networks.

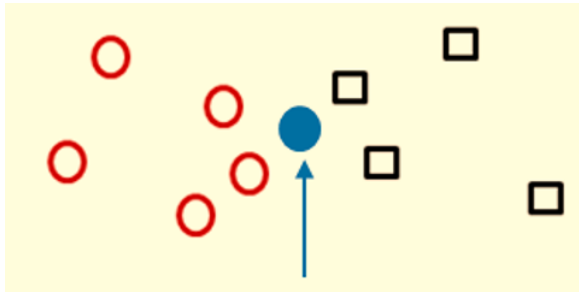
7.3.2.1. *K-nearest neighbors*

The kNN is an algorithm that can be used for both classification and regression. The principle of this model consists of choosing the k data closest to the studied point in order to predict its value. In classification or regression, the input will consist of the k closest training examples in a space.

In order to understand the functioning of this algorithm, we will take a small visual example. Below, we will show a training dataset, with two classes, circle and square. So, the input is bidimensional, and the target is to classify the data by shape.

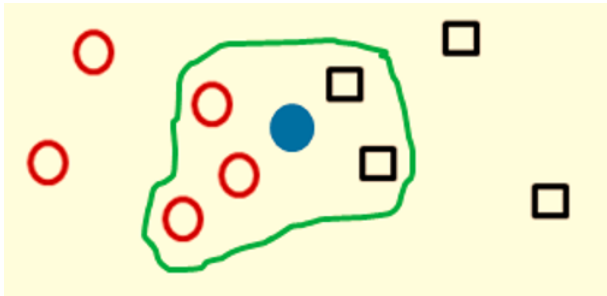


Now, if we have a new entry object whose class we want to predict, how could we do it?



The dark circle is a new entry

Well, we will simply look at the k closest neighbors to this point and see which class constitutes the majority of these points in order to deduce the class of the new entry.



For example, if we use the 5-NN, we can predict that the new entry belongs to the circle class since it has three circles and two squares in its entourage.

So, the principle of this algorithm is to classify a dataset in one of the categories by calculating the distance between it and each point of the training set. We choose the first k elements in order of distances, and therefore choose the dominant label among the k elements, which represents the category of the dataset element.

7.3.2.2. *Logistic regression*

This is a statistical method for performing binary classifications. It takes qualitative and/or ordinal predictors as input and measures the probability of the output value using the sigmoid function (see Figure 7.7). We can perform the multiclass classification (for example, classify a picture into three possibilities such as fruits, legumes and roses).

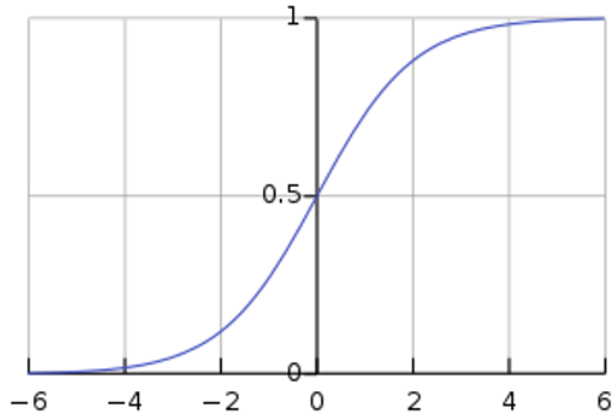


Figure 7.7. *Sigmoid function*

7.3.2.3. *Support Vector Machine*

SVM is also a binary classification algorithm. As shown in Figure 7.8, blue represents a class (non-spam mail for example) and red represents a spam. After tagging some words and concepts, the "signature" of the message can be injected into a classification algorithm to determine whether or not it is a spam.

Logistic regression can separate these two classes by defining the line in red. This method will opt to separate the two classes by the green line. Without going into details, and for mathematical considerations, the SVM will choose the clearest separation possible between the two classes (like the green line). This is why it is also called large margins classifier.

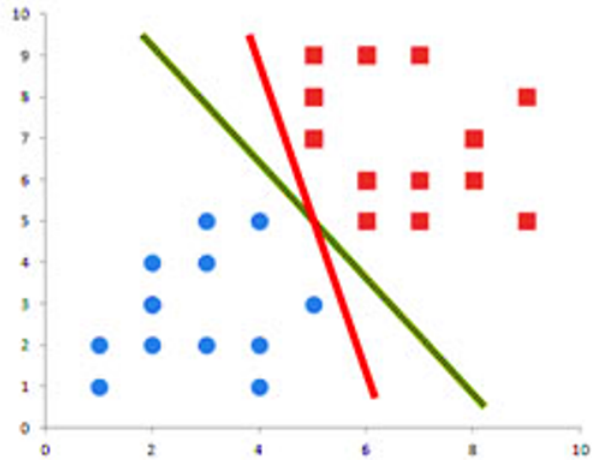


Figure 7.8. *Example of SVM. For a color version of this figure, see www.iste.co.uk/sedkaoui/data.zip*

7.3.2.4. *Naïve Bayes*

Naïve Bayes is a fairly intuitive classifier to understand. It assumes a strong (naïve) assumption. Indeed, it assumes that the variables are independent of each other. This simplifies the calculation of probabilities. Generally, Naïve Bayes is used for text classifications (based on the number of word occurrences).

Naïve Bayes classification is a machine learning method relying on Bayes' theorem:

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

It can be used for both binary and multiclass classification problems. The main point relies on the idea of treating each feature independently. The Naïve Bayes method evaluates the probability of each feature independently, regardless of any correlations, and makes the prediction based on Bayes' theorem. That is why this method is called "naïve"; in real-world problems, features often have some level of correlation between each other.

The advantages of using this method include its simplicity and ease of understanding. In addition, it performs well on the data sets with irrelevant features, since the probabilities contributing to the output are low. Therefore, they are not taken into account when making predictions.

Moreover, this algorithm usually results in good performance in terms of consumed resources, since it only needs to calculate the probabilities of the features and classes; there is no need to find any coefficients like in other algorithms. Its main drawback is that each feature is treated independently, although in most cases this cannot be true [BIS 06].

7.3.2.5. *Decision tree*

Another classification method is that of the decision tree. Decision trees are graph structures where each potential decision creates a new node, resulting in a tree-like graph [QUI 87]. The decision tree is an algorithm based on a graph model (the trees) to define the final decision. Each node has a condition, and the connections are based on this condition (true/false or yes/no or pass/fail). The further we descend into the tree, the more we combine the conditions. Figure 7.9 illustrates an example of this operation.

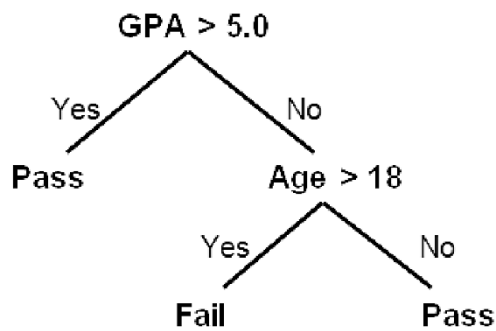


Figure 7.9. *Decision tree example*

A decision tree is built using a machine learning algorithm. Going from a set of predefined classes, the algorithm searches iteratively for the most different variables in the classified entities. Once this is identified, and the decision rules are determined, the dataset is segmented into several groups according to these rules. Data analysis is performed recursively on each subset until all key classification rules are identified.

7.3.2.6. *Random Forest*

Random Forest is one of the most popular machine learning algorithms. It requires almost no data preparation and modeling but usually produces accurate results. Random Forests are based on the decision trees described previously. More specifically, Random Forests are collections of decision trees, producing better prediction accuracy. That is why it is called a "forest" – it is basically a set of decision trees.

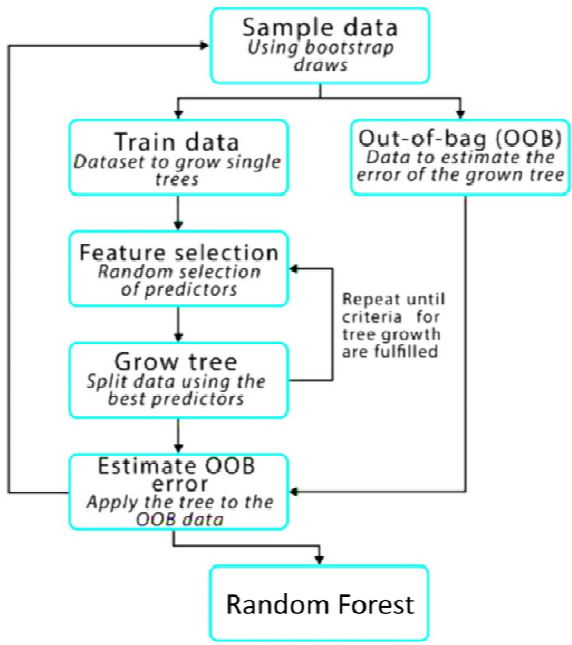


Figure 7.10. *Random Forest scheme*

7.3.2.7. *Neural networks*

Neural networks are inspired by the neurons of the human nervous system. They allow us to find complex patterns in the data. These neural networks learn a specific task based on the training data. Neural networks consist of nodes. In these networks, we find the input third (input layer) that will receive the input data.

The input layer will then propagate the data to hidden layers. Finally, the third party (output layer) produces the classification result. Each third of the neural network is a set of interconnections of the nodes of one third with those of the other thirds.

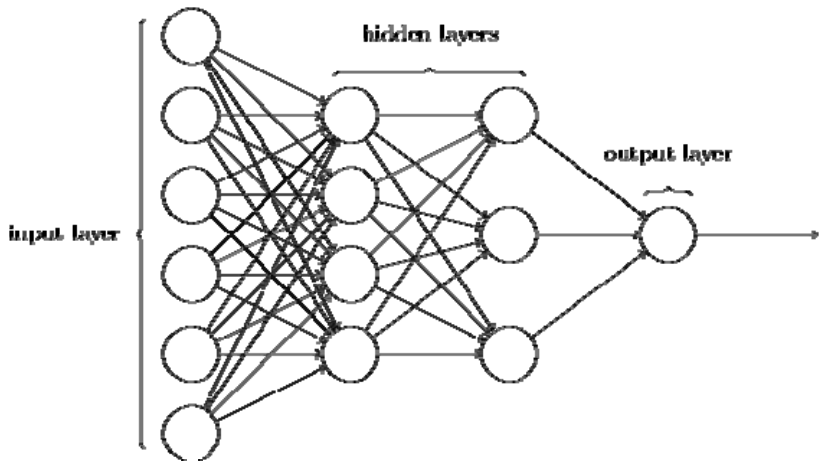


Figure 7.11. *Example of a neural network*

In these networks, the learning phase aims to converge the data parameters into an optimal classification. They require a lot of learning data and are not suitable for all problems, especially if the number of input parameters is too low.

In this case, the term *deep learning* refers to networks of juxtaposed neurons or consists of several layers. It draws upon, among other things, the latest advances in neuroscience and communication models of our nervous system. Some also associate it with modeling

that provides a higher level of data abstraction to produce better predictions.

Deep learning is particularly effective on the processing of images, sound and video. It is found in the fields of health, robotics, computer vision, etc.

Each of the algorithms cited above has its own mathematical and statistical properties. Depending on the training set and our features, we will opt for one or the other of these algorithms. However, the purpose is the same: *to predict to which class a datum belongs, for example whether a new email a spam or not.*

7.4. Clustering gathers data

ML is undoubtedly one of the major assets in understanding the challenges of society of today and tomorrow. Among the different components that make up this discipline, we will focus on one of the subdomains of application that characterize it: "clustering". This field covers diverse and varied subjects, and makes it possible to study the associated problems according to different perspectives. Indeed, we speak here of algorithmic objectivity.

7.4.1. *What good could it serve?*

Clustering aims to determine a segmentation of the studied population without *a priori* knowledge on the number of classes or "clusters", and to interpret *a posteriori* the clusters thus created. Here, humanity does not need to assist the machine in its different discovery typologies, since no target variable is provided to the algorithm during its learning phase.

Clustering algorithms are most often used for exploratory data analysis. This is, for example, to identify customers with similar behaviors (market segmentation), users who have similar uses, communities in social networks, etc.

The goal is to place the entities in a single large pool and form smaller groups that share similar characteristics; find the hidden patterns in the unlabeled data and separate it into clusters according to similarity.

Finding patterns in data with clustering algorithms seems to be an amazing work, but *what does it allow?*

At first glance, we might think that this method has little use in real-life applications. But, this is not the case, because the applications of clustering algorithm are numerous.

We can see it by wondering how Amazon (Amazon Web Services) recommends the right products, or how YouTube proposes to you several videos related to your expectations, or even how Netflix recommends good movies; all this is by applying a clustering algorithm.

The efficiency of applying a clustering algorithm can allow a significant increase in the turnover of an e-commerce site such as Amazon, for example.

Also, if we provide a set of pictures of animals without specifying what animals they are, then the algorithm will group together, for example, the pictures of dogs, of tigers and so on.

A cable TV that wants to determine the demographic distribution of network viewers can do so by creating clusters from available subscriber data and what they are watching. Another example can refer to the discovery of different customer groups inside the customer base of an online shop. Even a restaurant chain can group its customers according to the menus chosen by geographic location, and then modify its menus accordingly.

We can also mention the biomedical field as one of the extended fields of application of this algorithm, through, among other things, the grouping of differential genes according to their expression profile in a biological phenomenon over time.

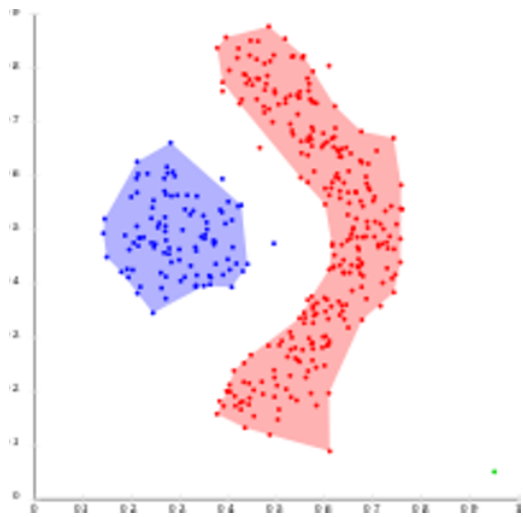
Also, for most music lovers, these algorithms can also be used, as already mentioned, for recommendation in order to distribute different music in clusters and to propose a song "similar" to that to which we just listened.

Some people on the web are planning to apply these methods to *Game of Thrones* characters in order to detect typologies of individuals, and who knows, perhaps scientifically determine who will be the real contenders for the iron throne.

The application examples are diverse and varied in the business context; we encounter this subdomain of ML through customer segmentation, a subject of considerable importance in the marketing community.

We can think more specifically of the detection of fraud, whether in public transport, as part of a complementary health reimbursement or regarding energy consumption.

So, the applications are numerous. For example, the points on the graph below can be considered similar if they are close in terms of distance.



In general, clustering algorithms examine a defined number of data characteristics and map each data entity to a corresponding point in a dimensional chart. The algorithms then seek to group the elements according to their relative proximity to each other in the graph.

7.4.2. *Principle of clustering algorithms*

When it comes to the non-labeled data, the ML algorithm will group these data by similarity. Since we are talking about similarity, we must also talk about "clustering". This is a family of unsupervised algorithms. Clustering algorithms fall into the unsupervised learning category. These make it possible to group together data that are similar.

This algorithm is an unsupervised learning task. The objective is to divide a set of objects, represented by inputs: $\{x_1, x_1, \ldots, x_1\}$, into a set of disjointed clusters: $\{\{x_{1,1}, x_{1,2}, \ldots, x_{1,n},\}, \{x_{2,1}, x_{2,2}, \ldots, x_{1,n}\}$, ……., $\{x_{n,1}, x_{n,2}, \ldots, x_{n,n}\}\}$, which contain objects similar to each other in some sense.

Clustering consists of grouping the data into homogeneous groups called classes or clusters, so that the elements within the same class are similar, and the elements belonging to two different classes are different. It is therefore necessary to define a measure of similarity between two elements of the data: the distance.

Each element can be defined by the values of its attributes, or what we call, from a mathematical point of view, a *vector*. The number of elements of this vector is the same for all elements and it is called the *vector dimension*, which is denoted by n.

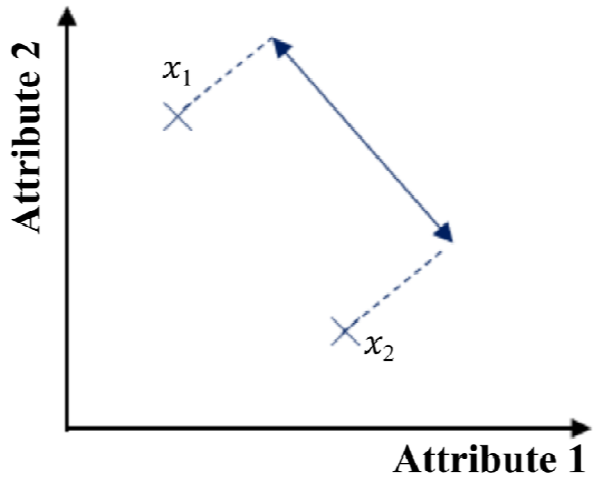
Given two vectors $V1$ and $V2$, we must define the distance between these two elements $d\ (V1, V2)$.

Typically, similarity between two objects is defined by Euclidean distance, Manhattan distance or Hamming distance.

7.4.2.1. *Euclidean distance*

This is the distance between two points. Considering two points p and q identified by their X and Y coordinates, we have:

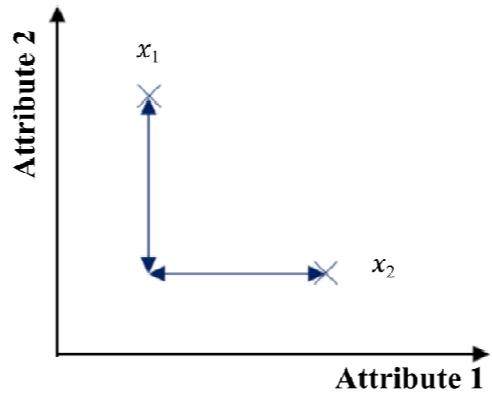
$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$



7.4.2.2. *Manhattan distance*

The name of this distance is inspired by the famous district of New York, consisting of many skyscrapers. It is impossible to go straight from one point to another and it is necessary to circumvent the buildings:

$$d(p,q) = \sum_{i=1}^{n}|p_i - q_i|$$



7.4.2.3. *Hamming distance*

This distance measures the similarity between two words by counting the number of characters that differ.

Perhaps you realize now that there are two important notions: "distance" and "similarity".

In Figure 7.12, you have an example of clustering. Points are sorted into two groups: Cluster 1 and Cluster 2. To calculate the similarity we chose the Euclidean distance as metric. This metric is the simplest and also the most intuitive. In Figure 7.12, the blue segment refers to the distance between the two points. Once we have calculated the distances between each point, the clustering algorithm automatically ranks the "near" (or "similar") points in the same group.

So, clustering refers to the methods of automatically grouping data that are most similar to one set of "clusters". A set of unsupervised algorithms can accomplish this task. They therefore automatically measure the similarity between the different data. Clustering algorithms therefore depend strongly on how we define this notion of similarity, which is often specific to the application domain.

The principle of the algorithm consists of assigning classes according to:

– minimizing the distance between the elements of the same cluster (intra-class distance);

– maximizing the distance between each cluster (interclass distance).

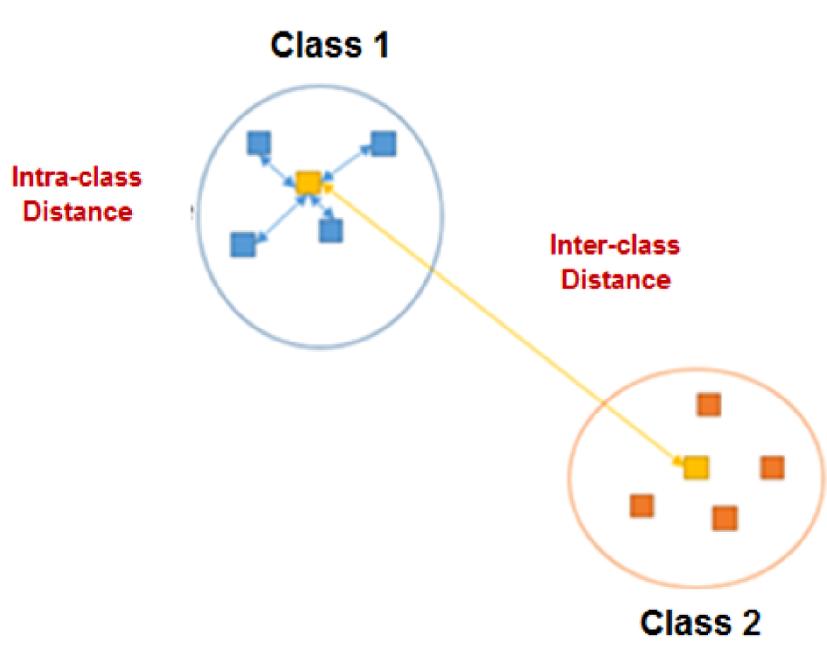


Figure 7.12. *Principle clustering algorithm*

However, to do this we have to try all possible combinations, and choose the solution with the minimum intraclass distances and the maximum interclass distances. To optimally assign classes to 27 elements, it would take a billion years for a processor of 3 GHz to achieve this task. This is why we use different algorithms to find the result that is as close as possible to the solution. This can be the case of K-means, for example, that need to be more detailed.

7.4.3. *Partitioning your data by using the K-means algorithm*

Let us suppose that you are looking to launch an advertising campaign and that you want to send a different advertising message depending on the target audience. First, you need to group the target population into groups. Individuals in each group will have a degree of similarity (age, gender, salary, etc.). That is what the K-means algorithm will do.

K-means is a type of clustering algorithm that is commonly used. This algorithm divides a set of data entities into groups, where k is the number of groups created. The algorithms refine the assignment of entities to different clusters by iteratively calculating the average midpoint or centroid of each cluster.

The centroids become the focal points of the iterations, which refine their locations in the plot and reassign the data entities to fit the new locations. An algorithm is repeated until the groupings are optimized and the centroids do not move anymore. The algorithm thus works as follows:

1) *initialization:* since the number of classes K is imposed, choose K points randomly to initially constitute the representatives of each class;

2) *then, for each point:*

– calculate the distances between this point and the classes' representatives: we begin by randomly choosing K centroids from our observations. Each point is then associated with the centroid of which it is closest;

– at this point assign the class from which its distance is minimal, thus forming K clusters;

– update the representatives of each class: we can now recalculate the centroid of each cluster (its center of gravity). Repeat the operation until the algorithm converges.

The illustration of the K-means algorithm execution result will help you to understand how it works.

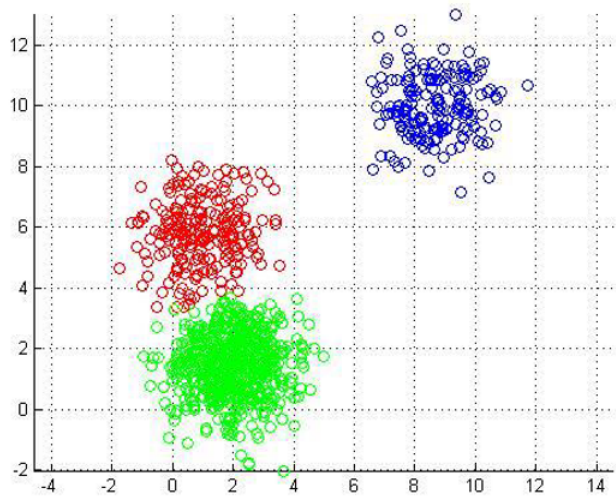


Figure 7.13. *Example of K-means clustering. For a color version of this figure, see www.iste.co.uk/sedkaoui/data.zip*

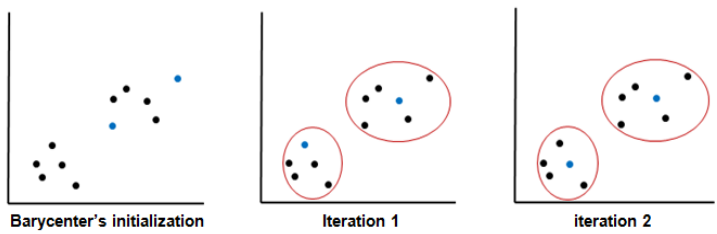
From Figure 7.13, we can recognize three clusters:

– cluster 1: in red;

– cluster 2: in blue;

– cluster 3: in green.

Data with one to three dimensions can be represented graphically. The others can only be apprehended mathematically. Here is a simplified representation of the iterative process that the algorithm will perform on two-dimensional data.



So, this algorithm solves the following problem: "given points and an integer K, the problem is to divide the points in K partitions so as to minimize a certain function".

Its speed compared to other algorithms makes it the most-used clustering algorithm. In practice, K does not correspond to the number of clusters that the algorithm will have to find and in which the elements will be stored, but rather to the number of centers (that is to say the central point of the cluster). It seems as if that is pretty much the same thing, but that allows us to go further.

Indeed, a cluster can be represented by a circle composed of a central point and a radius.

We seek to group N points. The algorithm will start with K center points that define the centers of each cluster to which the N points will be assigned at the end.

In the first step, the N points are associated with these different centers (initially specified or randomly selected); then, the next step is to recalculate the centers in relation to the average of all of the points of the cluster.

The first phase consists of finding the points that are associated with the cluster by calculating the distance between the center of the cluster and these points. In the second phase, the center of the cluster is recalculated relative to the average points of the cluster. We then start again until the centers stabilize, indicating that they arrived at optimal values to represent the N starting points.

However, this algorithm is limited by two elements:

– the number of clusters is defined by the user. It could very well be that our population has characteristics that make three clusters partition it better than two clusters. This can precede our K-means clustering with a *hierarchical classification* that will automatically define the best number of clusters to choose;

– the final clusters may depend on the random initialization of the centroids and thus propose different results for the same data. In the

complete algorithm, not presented here, I proposed a solution to initialize fictitious centers of gravity on a line crossing all of the dimensions of the data.

7.5. Conclusion

An ML algorithm is, as we have illustrated it in this chapter, a way of modeling phenomena in order to make strategic decisions. It is a very powerful tool when it is used well. An algorithm indicates how to combine and associate the data in order to obtain a response.

So, in order to carry out a process of data analytics using ML algorithms, it is best to consider an algorithm as a recipe, and data as ingredients, while the machine is like a mixer that supports many of the difficult tasks of the algorithm.

8

Applications and Examples

> "Some are born great, some achieve greatness, and some have greatness thrust upon them".
>
> Malvolio in Twelfth Night – Shakespeare

8.1. Introduction

Have you ever wondered how face or symbol detectors work on photos? The recommendation systems of Netflix, Amazon, Spotify and YouTube? Spam filters on your mailbox? All these systems and many others are part of the innumerable applications of ML algorithms. But which one should we choose: supervised or unsupervised?

This situation will be discussed in this chapter by highlighting other algorithms in the ML family. Also, in order to give a quick introduction to what is being done in ML applications and to trigger the reader's interest, we will introduce some examples of applications. This will allow you, in more detail, to gain more insight into the types and uses of ML algorithms.

8.2. Which algorithm to use?

If you ask a data scientist about which algorithm is best for analyzing such and such a problem, he will ask you to try several and see which one works best depending on your case.

Data Analytics and Big Data, First Edition. Soraya Sedkaoui.

So, it depends on:

– data quality;

– parameters that will be used;

– data source;

– execution time required;

– available parameters to influence the performance of the algorithm.

Therefore, for each type of question to be analyzed there is a specific group of methods or algorithms.

8.2.1. *Supervised or unsupervised algorithm: in which case do we use each one?*

ML algorithms are trained on annotated data (the training set) to build predictive models, or learners, which will enable us to predict the output of new unseen observations. It is called supervised because the learning process is done under the supervision of an output variable, in contrast with unsupervised learning where the response variable is not available.

In the case of supervised learning, the robustness of the algorithm will depend on the accuracy of its training. An algorithm learning supervised contents produces an internal map that allows its reuse to classify new data.

Take the example of an algorithm that detects faces: a user will have to show it what a face is and what it is not, so that the algorithm can learn and predict if the next picture refers to a face or not.

In summary, the algorithm learns from examples; in this case, the examples need to be labeled in order to ensure the effectiveness of its learning.

In the case of unsupervised learning, there is no need for the intervention of a human being because the algorithm will, by itself,

understand how to differentiate a face from a landscape by seeking their correlations. Since an algorithm cannot simply know what constitutes a face, the unsupervised method will classify the data into homogeneous groups called “clustering”.

In supervised learning, we have a set of observations. These observations (a dataset) contain a number of explanatory variables, and most often only one explained variable, or result. We want to be able to predict this result for future observations.

Supervised versus unsupervised	
Classification	Clustering
- Number of classes is known - Training - Use future data to classify	- Number of classes is unknown - No prior knowledge - Use to understand and exploit data

Table 8.1. *Supervised versus unsupervised algorithms*

The major difference when it comes to unsupervised learning is that there are no more examples. It is therefore no longer possible to evaluate the performance of the algorithm by comparing the result with observed values. The most popular unsupervised learning problems are clustering analytics.

Then, when you have a problem where you can precisely annotate for each observation the target you want to output, you can use supervised learning. Otherwise, it you want to better understand your dataset or identify interesting behaviors, you should use unsupervised learning.

The difference between the two algorithms is then fundamental. Supervised algorithms extract knowledge from a set of data containing input–output pairs. These pairs are already “known” in the sense that the outputs are defined *a priori*. The output value can be an expert-provided indication: for example, “true/false” or “pass/fail”.

These kinds of algorithms seek to define a compact representation of the input–output associations, from the data, via a prediction function.

However, unsupervised algorithms do not integrate the notion of input–output. All data are equivalent (we could say that there are only entries).

In this case, the algorithm seeks to organize the data into groups. Each group should have similar data and the different data should be in separate groups. Then, the learning is no longer based on an indication that can be previously provided by an expert but only from observable fluctuations in the data.

Table 8.2 illustrates some application examples for each one.

Question	Algorithm	Example
A or B or C…	Classification	To attract more customers, is it best to apply a $10 coupon on purchases that exceed $50 or a discount of 50%?
How much? How many?	Regression	How many (benefit) can the company achieve next year?
How are the data organized?	Clustering	What viewers like what types of movies?

Table 8.2. *Algorithm application examples*

A simple example would be helpful and can illustrate the principles of each one.

Imagine that a bank B has multiple information about its clients, such as gender, age, salary, operations, etc. Imagine that this bank is looking to anticipate their savings. To do this, B must adopt regression methods that make it possible, for example, to establish the link between the level of savings, the monthly salary and the automatic deductions.

Bank B also possesses some other information related to the typology of the clients corresponding to their distribution in different

categories: mass market, young archivers, etc. Bank B wants, in addition, to classify the new clients. It appeals to classification methods that allow bank B to predict which group they belong to, according to characteristics identified during a first appointment. Differentiated advantages will then be made for each group.

In both cases, the information to predict is identified in the form of a "label", that is, a predefined variable. However, unsupervised methods do not have this "label". These are the kind of algorithms that allow new structures to emerge.

We imagine the case of a new bank agency that does not have a classification model concerning these clients and which offers a new online service. This agency can discover unexpected things (new marketing tracks), for example, by identifying clients with a high level of connection on the website or who use mobile applications frequently (which mean that they respond less to requests for physical appointments).

8.2.2. *What about other ML algorithms?*

There are other questions which ML algorithms can answer, and which have not been mentioned in the previous chapter, such as "is this weird?".

To answer this question, you can use an algorithm called "anomaly detection".

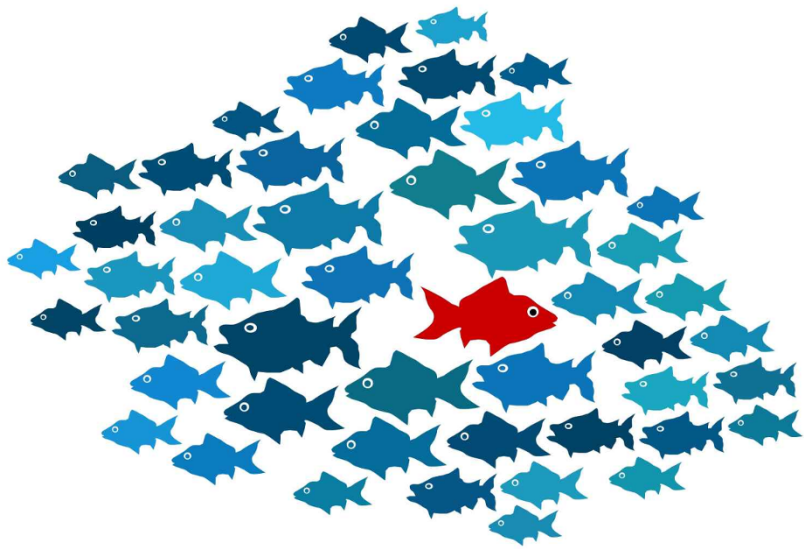
Anomaly detection is an ML algorithm for detecting abnormal patterns. For example, imagine that you received €1,500 into your bank account monthly and that one day you deposit €15,000 at once. The algorithm will detect this as an anomaly.

This algorithm is very useful for detecting fraud in banking transactions and intrusion detections.

If you have a credit card, you may benefit from anomaly detection. Your bank analyzes your purchase models to alert you in case of

fraud. For example, fees that are considered “strange” may be related to a purchase in a store where you buy an unusually expensive purchase.

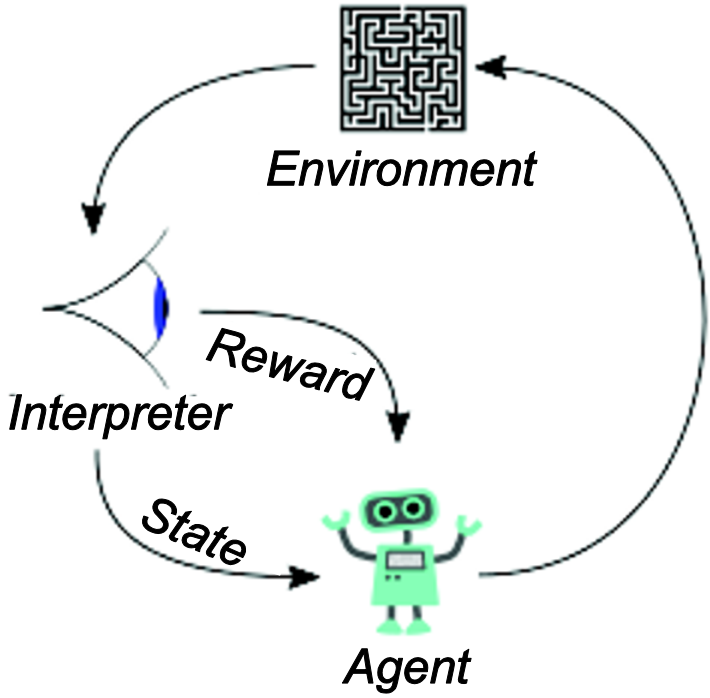
This algorithm signals unexpected or unusual events or behaviors. It gives guidance on where to look for problems.



The question “what should I do now?” uses an algorithm called a reinforcement learning algorithm.

This algorithm is inspired by how the brains of rats and humans respond to punishments and rewards. These algorithms learn on the basis of results in order to decide the next action.

Globally, this algorithm represents a solution perfectly suited to automated systems that must take a large number of small decisions without human instructions. For example, let’s think about a home temperature control system: should I adjust the temperature or leave it as it is?



Therefore, the questions which this algorithm answers concern the action to be carried out. It collects data as it goes along and learns from its tests and errors.

There is also *Transfer Learning* which aims to use the knowledge of a set of source tasks to not only influence learning but also improve performance on another target task. It consists of using the knowledge acquired to reapply it in another environment, for example, for two documents that are written in two different languages.

So, with this variety of algorithms, choosing the right one is a seemingly tedious process: there are many supervised and unsupervised machine learning algorithms, in addition to reinforcement and transfer learning algorithms, and each one approaches learning in a different way.

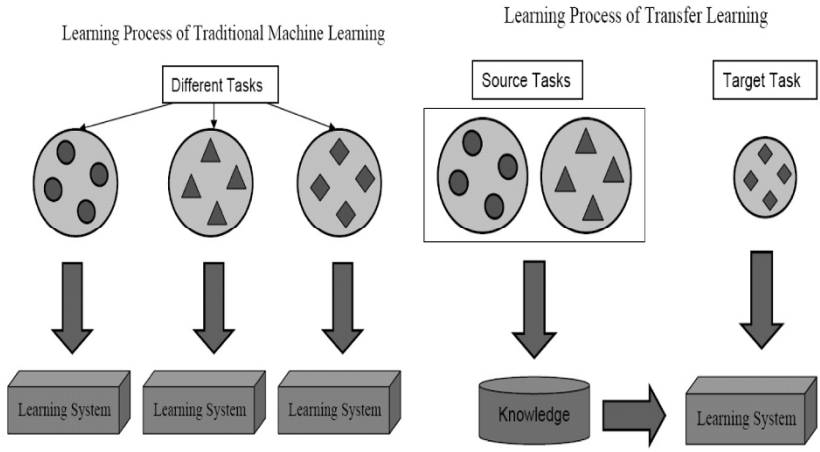


Figure 8.1. *Traditional learning versus transfer learning (source: https://www.computer.org/csdl/trans/tk/2010/10/ttk2010101345.html)*

There is no optimal or universal method. Determining the right algorithm to use is partly a question of trial and error. Even the most experienced data scientists cannot predict the proper functioning of an algorithm without performing prior tests. The choice of an algorithm also depends on the nature of your data and their volume, the information you want to extract and what you want to do with this information.

In summary, some tips that will help you in deciding are as follows:

1) opt for supervised learning if you want to train a model to make a forecast (for example the future value of a continuous variable such as the temperature or stock price) or a classification (for example identifying car brands appearing on video recordings of a webcam);

2) opt for unsupervised learning if you need to explore your data and want to drive a model to find good internal representations, for example by splitting data into clusters;

3) opt for reinforcement learning if you want to choose an action in response to each data point. Reinforcement learning is a common

approach in robotics, where the focus of sensor readings at a given moment is a data point and the algorithm must choose the next action of the robot. It is also suitable for IoT applications;

4) Opt for transfer learning to use a set of tasks to influence learning and improve performance in another task.

8.3. The duo big data/ML: examples of use

"ML algorithm" is an expression we have all heard in the media. Algorithms are a controversial subject. ML algorithms present a sort of real evolution. They make our programs smarter by allowing them to learn automatically from the data we provide. We are able to perform even better thanks to several applications: how do the algorithms work?

With the several applications existing today, you can experience how ML can generate added knowledge and how it turns ideas into business opportunities. ML revolutionizes businesses by utilizing the immensity of big data to draw unique observations and deductions, never envisaged, to better predict the next act of their clients: everything they have always wanted to know about their clients without ever daring to ask.

More and more data passes through the sieve of the algorithm: keywords entered in online searches, contents liked and disseminated on social networks, reading habits... All of this multitude of data, scattered on the web, form, when they are collected and analyzed, a complete portrait of each individual to better understand their tastes, desires, habits, etc. This mass of data is actually growing with the ever more intensive use of connected objects: computer, smartphone, smart car and so on.

In the end, this ultra-precise perception of behaviors makes it possible to refine and individualize the information of each client. Just think of Netflix, which suggests movies, or Amazon, who thinks it

knows the products we should buy. Several well-known examples seek to function according to everyone's preferences.

8.3.1. *Netflix: show me what you are looking at and I'll personalize what you like*

The strength of Netflix, like Facebook or Google, lies in the personalization of its recommendations. This is what has allowed the market of film and TV rental to reinvent itself. A few years ago, you had to go to a DVD store to watch a movie. A personalized recommendation for your next film could take place only if you knew the vendor.

Netflix is proud of its recommendation algorithms. The U.S. online video platform communicates regularly about how it is doing to make us consume the maximum content. "One last episode and I'll go to sleep!". Here is a sentence that should seem to you quite familiar and that illustrates well the "Netflix addiction".

Netflix's algorithm, coupled with machine learning (through the collection of data which evolves constantly), will allow the creation of diversity in the proposed visuals depending on what the user is likely to click on.

Let us take a closer look at how data are used and how these algorithms work. First, to access Netflix content, the user needs a subscription and therefore an individual account. It is the latter that makes the use of the data possible. Therefore, he is "tracked" as a unique individual with identifiable taste (what content for example), particular habits (day and time, how many times, etc.) but also sociological specificities and dozens of other criteria.

So, the important thing occurs when Netflix begins to consider other factors, such as when and how movies are viewed (device type, day of the week, schedule), how they are found, and so on. Netflix even begins to consider what content was recommended but not

clicked, using the failure of the algorithm as the source of information for the algorithm itself.

Once these data are collected, it remains only to cross and segment it to obtain groups of users (clients) with distinct characteristics. This is where the algorithm comes in. It will mix all of these data with the components specific to works in the Netflix catalog in order to offer movies and series which the user will almost certainly be interested in.

From analyzing the way a user is consuming and their passage from one series to another or from one movie to another, the algorithm will present a personalized "path" in adequacy with the user profile and observing their reaction.

Take the example of a user who has a preference for the genre of action movies according to his readings. One of the operating rules of Netflix will be to propose a movie of the same kind as the user preference (action) with another operating rule set according to the actors also playing in similar films. Another element of the algorithm personalizing the illustrations will be to highlight a known actor on a recommended movie.

Figure 8.2. *Profile personalization example for "Good Will Hunting" (source: Netflix)*

Thus, this work of recommendation and personalization by the illustrations proves a real challenge for the engineers of Netflix.

Many of them work continuously by screening the millions of personal data available.

To achieve their ends, engineers must collect as much data as possible to find the signals that one illustration is really better than another for a given subscriber, while avoiding too many tests by modifying images for the same content that risks disorientating users.

The challenge is also technical, both for designers who must create several illustrations per work (up to a few dozen) and for technical teams, who face the challenge of managing 20 million requests per second with a short time latency.

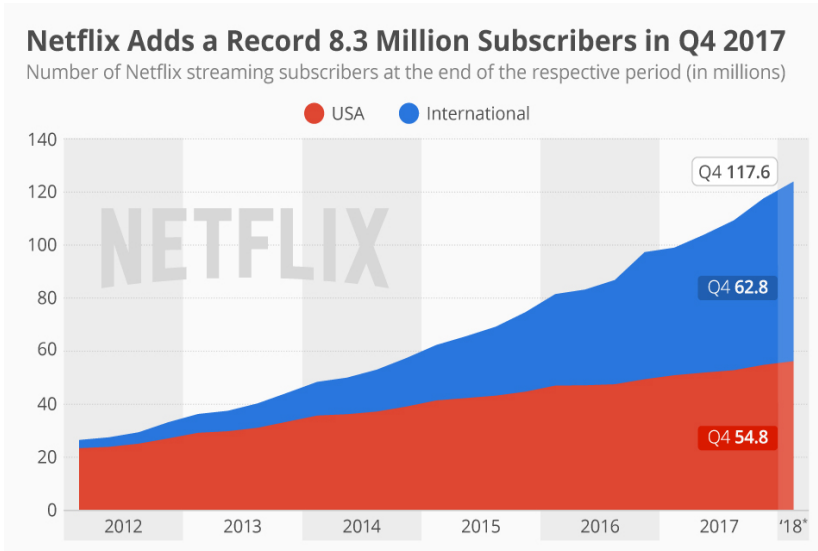


Figure 8.3. *Evolution of the number of Netflix subscribers (source: Statista Netflix, January 2018)*

Personalizing the platform based on the subscriber's usage history is at the heart of the Netflix's strategy, and the algorithms help automate the process as much as possible. Exported to more than 190 countries (Figure 8.3), Netflix continues its personalization work.

Figure 8.4. *Netflix net earnings (source: Statista Netflix, January 2018)*

There are more than 117 million subscribers in the world with around 6 million additional subscribers per quarter worldwide (and more than 8 million in the last quarter of 2017). Netflix net earnings continue to increase in the last quarter of 2017, as illustrated in Figure 8.4.

8.3.2. *Amazon: when AI comes into your everyday life*

More data above all provide a more precise target, one which takes into account your behavior on the web as a whole rather than on a single site. If you connect to an e-commerce site only for a specific product, it will only have a narrow view of your interests, so it will seek out data from news sites, social networks and others to better understand what interests you in terms of information and web searches.

Amazon uses navigation data from Facebook, Twitter, Google, etc. to refine targeting and better understand our interests. Amazon's case is also another important and interesting example of ML applications,

especially in the e-commerce field. For example, suppose we search for a product on Amazon today. When we come back another day, Amazon is able to offer us products related to our specific needs or our first research. This is because of ML algorithms that anticipate the evolution of our needs from our previous visits to Amazon.

Real-time marketing works on this principle, requiring a lot of data and, especially, the power of the cloud to retrieve live data collected by other sites to promote a bid or a personalized ad.

ML is not new to the global e-commerce giant. Already launched for over 20 years, Amazon has again recorded 22% growth in 2016. With Amazon Prime Now in June 2016 and Amazon Pantry in March 2017, innovations are set to fuel this growth.

Amazon Web Services (AWS) wants to take advantage of Amazon's experience and the strength of the group on the subject to simplify model training, accelerate learning phases and make ML accessible.

Latest: Amazon Echo. No need to move or go to the web to shop, a simple voice exchange with Alexa, a virtual assistant, is enough.

As the base installed in the heart of the intelligent home assistant Echo, the artificial intelligence *Alexa* is tending to become unavoidable. Alexa can be directly integrated into the operation of a washing machine, a refrigerator, a vacuum cleaner or a TV.

These appliances respond to more or less sophisticated voice commands in direct communication with the Alexa assistant, adapted here for the use of each product. The goal is to make such products easier to use: such as starting or stopping a wash cycle, checking and adjusting refrigerator temperatures, starting a recording on a connected TV, etc.

Amazon has already launched Look, a small connected camera able to analyze your appearance and tell you if you are well or badly dressed. Jeff Bezos' company goes further with an AI program based

on an algorithm able to design a garment. It is a sort of a fashion creative AI in a way.

In 2016, Amazon represented a larger market than most of the major players in the U.S. market combined. Look at Figure 8.5 to understand the situation.

Figure 8.5. *Amazon versus the major U.S. players (December 2016) (source: Visual capitalist)*

AWS does not just act on the infrastructure and development platform layers. It goes further by also offering turnkey cloud services for image recognition, video recognition and transcription of audio files into text, text translation and text analysis. They draw all their performances from ML. Developers can use them to quickly create face, object, or activity identification, tracking or content detection apps.

Amazon also develops concepts that everyone thought were still reserved for fantasy. This is the case, for example, with the impressive cashless grocery stores that are part of Amazon Go. Customers can simply use the displays and a complex system of chips, scales, cameras, AI and transmitters and receivers coupled to smartphones can track, validate, and pay for items without the user noticing.

This results in considerable time saving, a significant reduction in salary costs and an almost total elimination of the risks of theft. Other positive points could also be mentioned in terms of inventory

management, replenishment and sales analysis by period, by customer or by product, or the possible increase in expenses related to the simplicity of the concept and its fun aspect. The first store was opened in early 2017 in Seattle for Amazon employees exclusively as a test phase.

This is impressive; how does Amazon get this success? It is obviously the data. Since its creation in 1994, the company has adopted a culture largely driven by data. It is because of the knowledge given by the data that the customer gets the right product at the right time and is satisfied with the image of the Amazon logo: a smile.

8.3.3. *And more: proof that data are a source of creativity*

Globally, the success of GAFA (Google, Apple, Facebook, Amazon) and other companies such as Netflix, Airbnb, Tesla and Uber is based on the same power: value creation from data. Each of them has been able to identify data deposits and turn them into value. In the following, we will again demonstrate this importance by other significant big data applications.

8.3.3.1. *IBM: pattern recognition as part of diagnostics*

Like all fields, healthcare and the medical industry as a whole are gradually being turned upside down by big data. Through massive data collection and analysis, prevention, treatment, diagnostic and patient monitoring techniques are evolving rapidly.

For many years, healthcare, research and medical discoveries have relied on data collection and analysis. Health professionals then

sought to understand who gets sick, how and why. Today, thanks to the numerous sensors on smartphones and the increase in the quantity and quality of data, the possibilities of revolutionary discoveries are increasing.

Algorithms with ML skills are now more effective than humans in locating and diagnosing cancers. The potential of such AI is therefore immense and can make it possible to detect diseases earlier and to increase the chances of a successful treatment.

Programs like IBM Watson are used for pattern recognition in diagnostics. The Watson AI program is able to store, understand and explore large volumes of data through ML technologies, including deep learning.

This program is able to read press articles or research, tweets, novels or blog posts written in English and seven other languages because of solid training conducted by IBM researchers with natural language processing techniques.

IBM has been able to extract criteria for diagnosing heart failure from notes taken by physicians during consultations. They developed an ML algorithm that synthesizes the text using a technique called Natural Language Processing (NLP). In the same way as a cardiologist can read another doctor's notes and determine if a patient has heart failure, computers can now do the same thing.

Watson learns diagnostic methods in specialized partner institutions and IBM markets models in other clinics that do not have the same level of expertise. By reading the patient's genome, Watson can also make personalized recommendations. AI also uses its capabilities in image recognition, for example to read a radionuclide scan of a patient.

IBM therefore positions Watson as a health cloud, aiming to create a real ecosystem accessible in real time and around the world to cross-analyze (anonymized) patient data, a real "health infrastructure of the future".

Watson has been deployed in several hospitals and medical centers in recent years. Specifically, this system of "cognitive computing" ingests a body of contextual data belonging to a particular area of expertise selected by researchers and client companies.

8.3.3.2. *eBay*

Among the companies whose sustainability relies heavily on the use of big data technologies, we count eBay, the American giant of e-commerce. The company uses big data and ML to strengthen its business, but also to continue its development. Data, as the heart of the duo big data/ML, is today the most important asset of eBay. While this company has always been digital, today it embraces the latest big data technologies to enhance existing processes and create new experiences.

eBay's goal coupling big data and ML is to enable customization, merchandising and A/B testing of new features to enhance the user experience. ML makes it possible, for example, to improve the article recommendation system. This technology is also used for fraud detection and risk prediction for buyers and sellers.

To manage the different types and structures of data, as well as the speed required for analysis, the firm has moved from a traditional Data Warehouse to a "Data Lake". Inspired by web actors like Pinterest, ahead in the subject, eBay also invests in visual research.

Several months ago, eBay launched Image Search and Find It, two new features in this area. With Image Search, online shoppers make use of an image search bar to find similar products. With Find It, users will be able to share with eBay the URL of an image from social networks and the Marketplace app will list similar offers. These options can be used throughout the eBay catalog, almost 1.1 billion references.

8.3.3.3. *Spotify*

Since its launch in 2008 in Sweden, Spotify has raised the bar. The online music service is now available in many markets and has

over 100 million active users. With the end of 2016, Spotify launched its largest campaign ever. It has been deployed in 14 countries in total: France to begin with, then the United States, UK, Argentina, Australia, Brazil, Canada, Germany, Indonesia, Mexico, New Zealand, Philippines and Sweden for a second wave.

The idea is to dig into data to discover the strangest listening secrets of users. The Spotify systems use a network of artificial neurons as a learning method. Much like Netflix, Spotify uses ML to figure out users' likes and dislikes and provides them with a list of related tracks.

8.4. Conclusions

The aim of the ML process is to program machines to be able to learn from data and use examples to solve a given problem. Many successful applications of ML algorithms exist, from those that analyze data to predict customer behavior, to recognizing faces, recommending services or products filtering e-mail and so on.

We learned in this chapter to recognize a problem that uses ML, to deploy it with different methods and the ways to choose an algorithm among many others.

We have also investigated throughout the whole book such questions as: what is data? Why is it so important in today's context? How we can make sense of it in order to extract or generate value? What is the importance and the role of ML algorithms in big data? How can companies couple these two fields in order to make the data speak and reveal its secrets – secrets that can better conduct companies' strategies and be their guide in their decision-making process.

Congratulations! We have made it to the end of this book and with it its main purpose which is to investigate, explore and describe it approaches and methods to facilitate data understanding through analytics solutions and ML algorithms based on its principles, concepts and applications.

We hope you had fun reading and you had acquired some basic insights related to the big data universe. Now, try to work your way through the examples and applications to choose your path. And know that “there’s a difference between knowing the path and walking the path” (Morpheus, The Matrix). So, what you really need is practice, practice and practice!

Bibliography

[ACK 89] ACKOFF R.L., "From data to wisdom", *Journal of Applied Systems Analysis*, vol. 15, pp. 3–9, 1989.

[ACK 96] ACKOFF R.L., "On learning and the systems that facilitate it", *Center for Quality of Management Journal*, vol. 5, no. 2, pp. 27–35, 1996.

[AGR 11] AGRAWAL D., BERNSTEIN P., BERTINO E., "Challenges and opportunities with big data: A community white paper developed by leading researchers across the United States", *Tech. Rep.*, 2011.

[AKE 14] AKERKAR R., *Big Data Computing*, CRC Press, Boca Raton, 2014.

[BEC 86] BECKER H.B., "Can users really absorb data at today's rates? Tomorrow's?", *Data Communications*, vol. 15, no. 8, pp. 177–193, 1986.

[BER 13] BERMAN J.J., *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information*, Elsevier, Amsterdam, The Netherlands, 2013.

[BEY 12] BEYER M.A, LANEY D., The Importance of 'Big Data': A Definition, Gartner, 2012.

[BIS 06] BISHOP C., *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[BOR 16] BORT J., How IBM Watson saved the life of a woman dying from cancer, exec says. Retrieved November 25, 2017, available at: http://uk.businessinsider.com/how-ibm-watson-helped-cure-a-womans-cancer-2016-12?r=US&IR=T, 7 December 2016.

Data Analytics and Big Data, First Edition. Soraya Sedkaoui.

[BRY 11] BRYNJOLFSSON E., MCAFEE A., *Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy*, Digital Frontier Press, Lexington, MA, 2011.

[CHA 74] CHAMBERLIN D.D., BOYCE R.F., SEQUEL: A structured English query language. *Proc. 1974 ACM SIGFIDET Workshop*, Ann Arbor, Michigan, pp. 249–264, April 1974.

[CHE 12] CHEN H., CHIANG R.H., STOREY V.C., "Business intelligence and analytics: From big data to big impact", *MIS Quarterly*, vol. 36, no. 4, 2012.

[CHO 11] CHOI H., VARIAN H., Predicting the present with Google Trends, Google Research, 2011.

[COD 70] CODD E.F., "A relational model of data for large shared data banks", *Comm*, *ACM*, vol. 13, no. 6, pp. 377–387, June 1970.

[COO 12] COOPER A., "What is analytics? Definition and essential characteristics", *CETIS Analytics Series*, vol. 1, no. 5, pp. 1–10, 2012.

[CUK 13a] CUKIER K., MAYER-SCHONBERGER V., *Big Data: A Revolution That Will Transform How We Live, Work and Think*, Houghton Mifflin Harcourt, Boston, MA, 2013.

[CUK 13b] CUKIER K., MAYER-SCHOENBERGER V., "The Rise of Big Data", *Foreign Affairs*, May/June, vol. 92, no. 3, pp. 28–40, 2013.

[DAV 07] DAVENPORT T.H., HARRIS J.G., "Computing analytics: the new science of winning", Harvard Business School Review Press, Boston, MA, 2007.

[DAV 13] DAVENPORT T.H., KIM J., "Keeping Up with the Quants: Your Guide to Understanding and Using Analytics", Harvard Business Review Press, Boston, MA, 2013.

[DAV 14] DAVENPORT T.H., "Big data at work: Dispelling the myths, uncovering the opportunities", Harvard Business Review Press, Boston, MA, 2014.

[DEL 13] DELEN D., DEMIRKAN H., "Data, information and analytics as services", *Decision Support Systems*, vol. 55, no. 1, pp. 359–363, 2013.

[DIE 14] DIETRICH B.L., PLACHY E.C., NORTON M.F., *Analytics across the Enterprise, How IBM realizes Business Value from Big Data and Analytics*, IBM Press, New York, 2014.

[DRE 17] DRESNER, Adivsory Big Data Analytics Market Study, 3rd annual report, 2017.

[ECK 07] ECKERSON W., Predictive analytics: Extending the Value of Your Data Warehousing Investment, TDWI Best Practices Report, vol. 1, pp. 1–36, 2007.

[EMC 14] EMC., IDC, The digital universe of opportunities: Rich data and the increasing value of the Internet of Things, EMC, IDC, 2014.

[FOS 17] FOSTER I., GHANI R., JARMIN R.S. *et al*., *Big Data and Social Science*, CRC Press, Boca Raton, 2017.

[FRA 08] FRANKEL F., REID R., "Big Data: distilling meaning from data", *Nature*, vol. 455, pp. 30–30, 2008.

[FRI 16] FRIZZO-BARKER J., CHOW-WHITE P.A., MOZAFARI M. *et al*., "An empirical study of the rise of big data in business scholarship", *International Journal of Information Management*, vol. 36, no. 3, pp. 403–413, 2016.

[FUJ 12] FUJIMAKI R., MORINAGA S., "The Most Advanced Data Mining of the Big Data Era", *Advanced technologies to support big data processing*, vol. 7, no. 2, 2012.

[GAN 11] GANTZ B.J., REINSEL D., Extracting value from chaos state of the universe. Retrieved from: http://www.emc.com/collateral/analyst-reports/idc-extracting-valuefrom-chaos-ar.pdf, 2011.

[GAN 15] GANDOMI A., HAIDER M., "Beyond the hype: big data concepts, methods, and analytics", *International Journal of information management*, vol. 35, no. 2, pp. 137–144, 2015.

[GAR 13] GARTNER, Gartner IT Glossary: Big Data, Retrieved from: http://www.gartner.com/it-glossary/ big-data/, 2013.

[GAR 17] GARTNER, Newsroom, Stamford, Conn., Retrieved from: https://www. gartner.com/newsroom/id/3568917, 12 January 2017.

[GUT 15] GUTIERREZ D., *Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R.,* Technics Publications, New Jersey, USA, 2015.

[HAR 14] HARFORD T., "Big Data: A Big Mistake", *Significance*, vol. 11, no. 5, pp. 14–19, 2014.

[HAZ 14] HAZEN B.T., BOONE C.A., EZELL J.D. *et al*., "Data quality for data science, predictive analytics and Big Data in supply chain management: An introduction to the problem and suggestions for research and applications", *Int. J. Production Econ*., vol. 154, pp. 72–80, 2014.

[INS 17] INSTITUTE OF INTERNATIONAL FINANCE, Deploying RegTech Against Financial Crime, Report of the IIF RegTech Working Group, March, 2017.

[INU 14] INUKOLLU V.N., KESHAMONI D.D., KANG T. *et al*., "Factors influencing quality of mobile apps: role of mobile app development life cycle", *International Journal of Software Engineering & Applications*, vol. 5, no. 5, pp. 15–34, 2014.

[JAG 12] JAGADISH S.V.K., SEPTININGSIH E.M., KOHLI A. *et al*., "Genetic advances in adapting rice to a rapidly changing climate", *J Agron Crop Sci*, vol. 198, no. 5, pp. 360–373, 2012.

[JOR 13] JORDAN M.I., "On statistics, computation and scalability", *Bernoulli*, vol. 19, no. 4, pp. 1378–1390, 2013.

[KAT 13] KATAL A., WAZID M., GOUDAR R.H., "Big Data: Issues, Challenges, Tools and Good Practices", *IEEE Spectrum*, pp. 404–409, 2013.

[KIM 11] KIMBALL R., ROSS M., *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*, John Wiley & Sons, 2011.

[KLE 17] KLEIN A., Hard Drive Cost Per Gigabyte, Backblaze, July, 2017.

[KRI 10] KRIVKO M., "A hybrid model for plastic card fraud detection systems", *Expert Systems with Applications*, vol. 37, no. 8, pp. 6070–6076, 2010.

[KUD 14] KUDYBA S., "Information Creation through Analytics", in S. KUDYBA (ed.), *Big Data, Mining and Analytics*, *Components of Strategic Decision Making,* CRC Press, Boca Raton, pp. 17–48, 2014.

[LAB 12] LABRINIDIS A., JAGADISH H.V., "Challenges and opportunities with big data", *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 2032–2033, 2012.

[LIU 17] LIU G., YANG H., "Self-organizing network for variable clustering", *Annals of Operations Research*, 2017.

[MAN 11] MANYIKA J., CHUI M., BROWN B. *et al*., Big data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, Washington DC, USA, 2011.

[MAR 01] MARTINET B., MARTI Y.-M., *L'intelligence économique : Comment donner de la valeur concurrentielle à l'information*, Editions d'Organisation, Paris, France, 2001.

[MAR 16] MARR B., *Big Data in Practice (Use Cases): How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results*, John Wiley & Sons, 2016.

[MCK 11] MCKINSEY GLOBAL INSTITUTE, Big data: The next frontier for innovation, competition, and productivity, New York, USA, 2011.

[MCK 13] MCKINSEY GLOBAL INSTITUTE, Big Data, Analytics and the Future of Marketing and Sales, New York, USA, 2013.

[MCK 16] MCKINSEY GLOBAL INSTITUTE, The age of analytics: Competing in a Data-driven world, New York, USA, December 2016.

[MIT 97] MITCHELL T.M., *Machine Learning*, McGraw-Hill, 1997.

[MIT 02] MITRA S., PAL S.K., MITRA P., "Data mining in soft computing framework: A survey", *IEEE Transactions on Neural Networks*, vol. 13, pp. 3–14, 2002.

[MOH 12] MOHRI M., ROSTAMIZADEH A., TALWALKAR A., *Foundations of Machine Learning*, MIT Press, 2012.

[MOO 15] MOORTHY J., LAHIRI R., BISWAS N. *et al.*, "Big data: Prospects and challenges", *Sage*, India, 2015.

[MOR 15] MORABITO V., *Big Data and Analytics: Strategic and Organizational Impacts*, Springer International Publishing, Switzerland, 2015.

[MUH 13] MUHTAROGLU F.C.P., DEMIR S., OBALI M. *et al.*, "Business model canvas perspective on big data applications", *2013 Proceedings: IEEE International Conference on Big Data*, pp. 32–37, 2013.

[MUI 14] MUI Y., The weird Google searches about unemployment and what they say about the economy, available at: http://www.washingtonpost.com/blogs/wonkblog/wp/2014/05/30/the-weird-google-searchesof-the-unemployed-and-what-they-say-about-the-economy/, 2014.

[NYC 07] NYCE C., Predictive Analytics White Paper, American Institute for CPCU, pp. 9–10, 2007.

[OHL 13] OHLHORST F., *Big Data Analytics: Turning Big Data into Big Money*, John Wiley & Sons, Inc, New Jersey, USA, 2013.

[PAT 15] PATTY J., PENN E.M., "Analyzing Big Data: Social Choice and Measurement", *PS: Political Science and Politics*, vol. 48, no. 1, 2015.

[PIE 15] PIEGORSCH W.W., *Statistical Data Analytics*, John Wiley & Sons, New York, 2015.

[POP 12] POPOVIČ A., HACKNEY R., COELHO P.S. *et al.*, "Towards business intelligence systems success: effects of maturity and culture on analytical decision making", *Decis, Support Syst*, vol. 54, pp. 729–739, 2012.

[POR 85] PORTER M.E., *Competitive Advantage: Creating and Sustaining Superior Performance*, Free Press, New York, USA, 1985.

[POR 15] PORTER M., HEPPELMANN J.-E., "How smart, connected products are transforming Competition", *Harvard Business Review*, Boston, MA, November, 2015.

[QUI 87] QUINLAN J.R., "Simplifying decision trees", *International Journal of ManMachine Studies*, vol. 27, no. 3, pp. 221–234, 1987.

[QUO 12] QUOC V.L., RANZATO M.A., MONGA R. *et al*., " Building high-level features using large scale unsupervised learning", *ICML*, pp. 507–514, 2012.

[RAY 95] RAYPORT J.F., SVIOKLA J.J., "Exploiting the Virtual Value Chain", *Harvard Business Review*, vol. 73, no. 6, pp.75–85, 1995.

[REI 17] REINSEL D., GANTZ J., RYDNING J., Data Age 2025: The Evolution of Data to Life-Critical, IDC White Paper, April 2017.

[SAS 17] SAS, Predictive Analytics: What it is and why it matters, available at: https://www.sas.com/en_us/insights/analytics/predictive-analytics.html, 2017.

[SCH 14] SCHLEGEL G.L., "Utilizing Big Data and Predictive Analytics to Manage Supply Chain Risk", *Journal of Business Forecasting*, vol. 33, no. 4, pp. 11–17, 2014.

[SED 16] SEDKAOUI S., MONINO J.L., *Big Data, Open Data and Data Development*, ISTE Ltd, London and John Wiley & Sons, New York, USA, 2016.

[SED 17] SEDKAOUI S., "The Internet, Data Analytics and Big Data", in GOTTINGER H.W (ed.), *Internet Economics: Models, Mechanisms and Management*, eBook Bentham Science Publishers, UAE, Sharjah, 2017.

[SHO 12] SHOCKLEY R., Analytics: The real-world use of big data, IBM, 2012.

[SHM 11] SHMUELI G,. KOPPIUS O.R., "Predictive analytics in information systems research", *MIS Quarterly*, vol. 35, no. 3, pp. 553–572, 2011.

[SHR 13] SHROFF G., *The Intelligent Web, Search, Smart Algorithms and Big Data*, Oxford University Press, Oxford, 2013.

[SIE 13] SIEGEL E., *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*, John Wiley & Sons, Hoboken, NJ, 2013.

[SIE 16] SIEGEL E., *Predictive Analytics*, John Wiley & Sons, New York, 2016.

[SIM 77] SIMON H.A., *The New Science of Management Decision*, Prentice Hall, Englewood Cliffs, NJ, 1977.

[SIM 97] SIMON H.A., *Administrative Behavior: A Study of Decision-making Processes in Administrative Organizations*, Free Press, New York, USA, 1997.

[SOL 17] SOLDATOS J., *Building Blocks for IoT Analytics Internet-of-Things Analytics*, River Publishers Series in Signal, Image and Speech Processing, 2017.

[STA 14] STARKEY C.M., SCHNIEDERJANS M.J., SCHNIEDERJANS D.G., *Business Analytics Principles, Concepts, and Applications: What, Why, and How*, Pearson Inc, New Jersey, USA, 2014.

[STU 11] STUBBS E., *The Value of Business Analytics*, John Wiley & Sons, Hoboken, NJ, 2011.

[SUG 15] SUGIYAMA M., *Introduction to Statistical Machine Learning*, Morgan Kaufmann, 2015.

[SUM 06] SUMATHI S., SIVANANDAM S.N., *Introduction to Data Mining and its Application*, Springer, New York, USA, 2006.

[TAY 80] TAYLOR R.S., "Value-added aspects of the information process", *Communicating Information: Proceedings of the 43rd ASIS Annual Meeting*, Anaheim, CA, vol. 17, pp. 5–10, October 1980.

[THE 10] THE ECONOMIST, Data, data everywhere. A special report on managing information, 27th February 2010.

[THE 17] THEOBALD O., *Data Analytics for Absolute Beginners*, Scatterplot Press, Kindle edition, 2017.

[TOM 16] TOMAR G.S., CHAUDHARI N.S., BHADORIA R.S. *et al*., (eds), *The Human Element of Big Data: Issues, Analytics, and Performance*, CRC Press, Boca Raton, 2016.

[TUK 77] TUKEY J.W., *Exploratory Data Analysis*, Addison-Wesley, Reading, MA, 1977.

[TUR 11] TURBAN E., LIANG T.P., WU S.P., "A framework for adopting collaboration 2.0 tools for virtual group decision making", *Group Decision and Negotiation*, vol. 20, no. 2, pp. 137–154, 2011.

[VAG 14] VAGATA P., WILFONG K., Scaling the Facebook data warehouse to 300 PB, Facebook, available at: https://goo.gl/cMpqJM, 2014.

[VAN 12] VAN BARNEVELD A., ARNOLD K.E., CAMPBELL J.P., "Analytics in higher education: Establishing a common language", *EDUCAUSE*, vol. 1, no. 1, pp. 1–11, 2012.

[VAS 15] VASARHELYI M.A., KOGAN A., TUTTLE B.M., "Big Data in Accounting: An Overview", *Accounting Horizons*, vol. 29, no. 2, pp. 381–396, 2015.

[WAL 13] WALLER M.A., FAWCETT S.E., "Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management", *Journal of Business Logistics*, vol. 34, pp. 77–84, 2013.

[WAL 16] WALWEI U., Digitalization and structural labor market problems: The case of Germany, ILO Research Paper, vol. 17, pp. 1–31, 2016.

[WAN 15] WANG X., GUO F., HELLER K.A. *et al.*, "Parallelizing MCMC with Random Partition Trees", *NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems,* arXiv preprint, pp. 1506–03164, 2015.

[WAR 15] WARREN J., DONALD J., MOFFITT K. C. *et al.*, "How Big Data Will Change Accounting", *Accounting Horizons*, vol. 29, no. 2, pp. 397–407, 2015.

[WAT 16] WATT J., BORHANI R., KATSAGGELOS A., *Machine Learning Refined: Foundations, Algorithms and Applications*, Cambridge University Press, 2016.

[WIT 16] WITTEN I.H., FRANK E., HALL M.A. *et al.*, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2016.

[WOR 16] World Economic Forum Global Information Technology Report, 2016.

[ZHA 17] ZHANG A., *Data Analytics: Practical Guide to Leveraging the Power of Algorithms, Data Science, Data Mining, Statistics, Big Data, And Predictive Analysis to Improve Business, Work, and Life*, Kindle edition, 2017.

[ZIC 14] ZICARI R.V., "Big Data: Challenges and Opportunities", in AKERKA R. (ed.), *Big Data Computing*, CRC Press, Boca Raton, 2014.

Index

Data Analytics and Big Data, First Edition. Soraya Sedkaoui.

E

F, G

H, I, J

K, L

M

N

O, P

Q, R

S

T

U, V

W, Y, Z

Other titles from

in

Computer Engineering

2018

ANDRO Mathieu
Digital Libraries and Crowdsourcing
(Digital Tools and Uses Set – Volume 5)

ARNALDI Bruno, PASCAL Guitton, GUILLAUME Moreau
Virtual Reality and Augmented Reality: Myths and Realities

HOMAYOUNI S. Mahdi, FONTES Dalila B.M.M.
Metaheuristics for Maritime Operations
(Optimization Heuristics Set – Volume 1)

2017

BENMAMMAR Badr
Concurrent, Real-Time and Distributed Programming in Java

HÉLIODORE Frédéric, NAKIB Amir, ISMAIL Boussaad, OUCHRAA Salma, SCHMITT Laurent
Metaheuristics for Intelligent Electrical Networks
(Metaheuristics Set – Volume 10)

MA Haiping, SIMON Dan

Evolutionary Computation with Biogeography-based Optimization
(Metaheuristics Set – Volume 8)

PÉTROWSKI Alain, BEN-HAMIDA Sana
Evolutionary Algorithms
(Metaheuristics Set – Volume 9)

PAI G A Vijayalakshmi
Metaheuristics for Portfolio Optimization
(Metaheuristics Set – Volume 11)

2016

BLUM Christian, FESTA Paola
Metaheuristics for String Problems in Bio-informatics
(Metaheuristics Set – Volume 6)

DEROUSSI Laurent
Metaheuristics for Logistics
(Metaheuristics Set – Volume 4)

DHAENENS Clarisse and JOURDAN Laetitia
Metaheuristics for Big Data
(Metaheuristics Set – Volume 5)

LABADIE Nacima, PRINS Christian, PRODHON Caroline
Metaheuristics for Vehicle Routing Problems
(Metaheuristics Set – Volume 3)

LEROY Laure
Eyestrain Reduction in Stereoscopy

LUTTON Evelyne, PERROT Nathalie, TONDA Albert
Evolutionary Algorithms for Food Science and Technology
(Metaheuristics Set – Volume 7)

MAGOULÈS Frédéric, ZHAO Hai-Xiang
Data Mining and Machine Learning in Building Energy Analysis

RIGO Michel
Advanced Graph Theory and Combinatorics

2015

BARBIER Franck, RECOUSSINE Jean-Luc
COBOL Software Modernization: From Principles to Implementation with the BLU AGE® Method

CHEN Ken
Performance Evaluation by Simulation and Analysis with Applications to Computer Networks

CLERC Maurice
Guided Randomness in Optimization
(Metaheuristics Set – Volume 1)

DURAND Nicolas, GIANAZZA David, GOTTELAND Jean-Baptiste, ALLIOT Jean-Marc
Metaheuristics for Air Traffic Management
(Metaheuristics Set – Volume 2)

MAGOULÈS Frédéric, ROUX François-Xavier, HOUZEAUX Guillaume
Parallel Scientific Computing

MUNEESAWANG Paisarn, YAMMEN Suchart
Visual Inspection Technology in the Hard Disk Drive Industry

2014

BOULANGER Jean-Louis
Formal Methods Applied to Industrial Complex Systems

BOULANGER Jean-Louis
Formal Methods Applied to Complex Systems:
Implementation of the B Method

GARDI Frédéric, BENOIST Thierry, DARLAY Julien, ESTELLON Bertrand, MEGEL Romain
Mathematical Programming Solver based on Local Search

KRICHEN Saoussen, CHAOUACHI Jouhaina
Graph-related Optimization and Decision Support Systems

LARRIEU Nicolas, VARET Antoine
Rapid Prototyping of Software for Avionics Systems: Model-oriented Approaches for Complex Systems Certification

OUSSALAH Mourad Chabane
Software Architecture 1
Software Architecture 2

PASCHOS Vangelis Th
Combinatorial Optimization – 3-volume series, 2nd Edition
Concepts of Combinatorial Optimization – Volume 1, 2nd Edition
Problems and New Approaches – Volume 2, 2nd Edition
Applications of Combinatorial Optimization – Volume 3, 2nd Edition

QUESNEL Flavien
Scheduling of Large-scale Virtualized Infrastructures: Toward Cooperative Management

RIGO Michel
Formal Languages, Automata and Numeration Systems 1: Introduction to Combinatorics on Words
Formal Languages, Automata and Numeration Systems 2: Applications to Recognizability and Decidability

SAINT-DIZIER Patrick
Musical Rhetoric: Foundations and Annotation Schemes

TOUATI Sid, DE DINECHIN Benoit
Advanced Backend Optimization

2013

ANDRÉ Etienne, SOULAT Romain
The Inverse Method: Parametric Verification of Real-time Embedded Systems

BOULANGER Jean-Louis
Safety Management for Software-based Equipment

DELAHAYE Daniel, PUECHMOREL Stéphane
Modeling and Optimization of Air Traffic

FRANCOPOULO Gil
LMF — Lexical Markup Framework

GHÉDIRA Khaled
Constraint Satisfaction Problems

ROCHANGE Christine, UHRIG Sascha, SAINRAT Pascal
Time-Predictable Architectures

WAHBI Mohamed
Algorithms and Ordering Heuristics for Distributed Constraint Satisfaction Problems

ZELM Martin *et al.*
Enterprise Interoperability

2012

ARBOLEDA Hugo, ROYER Jean-Claude
Model-Driven and Software Product Line Engineering

BLANCHET Gérard, DUPOUY Bertrand
Computer Architecture

BOULANGER Jean-Louis
Industrial Use of Formal Methods: Formal Verification

BOULANGER Jean-Louis
Formal Method: Industrial Use from Model to the Code

CALVARY Gaëlle, DELOT Thierry, SEDES Florence, TIGLI Jean-Yves
Computer Science and Ambient Intelligence

MAHOUT Vincent
Assembly Language Programming: ARM Cortex-M3 2.0: Organization, Innovation and Territory

MARLET Renaud
Program Specialization

SOTO Maria, SEVAUX Marc, ROSSI André, LAURENT Johann
Memory Allocation Problems in Embedded Systems: Optimization Methods

2011

BICHOT Charles-Edmond, SIARRY Patrick
Graph Partitioning

BOULANGER Jean-Louis
Static Analysis of Software: The Abstract Interpretation

CAFERRA Ricardo
Logic for Computer Science and Artificial Intelligence

HOMES Bernard
Fundamentals of Software Testing

KORDON Fabrice, HADDAD Serge, PAUTET Laurent, PETRUCCI Laure
Distributed Systems: Design and Algorithms

KORDON Fabrice, HADDAD Serge, PAUTET Laurent, PETRUCCI Laure
Models and Analysis in Distributed Systems

LORCA Xavier
Tree-based Graph Partitioning Constraint

TRUCHET Charlotte, ASSAYAG Gerard
Constraint Programming in Music

VICAT-BLANC PRIMET Pascale *et al.*
Computing Networks: From Cluster to Cloud Computing

2010

AUDIBERT Pierre
Mathematics for Informatics and Computer Science

BABAU Jean-Philippe *et al.*
Model Driven Engineering for Distributed Real-Time Embedded Systems 2009

BOULANGER Jean-Louis
Safety of Computer Architectures

MONMARCHE Nicolas *et al.*
Artificial Ants

PANETTO Hervé, BOUDJLIDA Nacer
Interoperability for Enterprise Software and Applications 2010

SIGAUD Olivier *et al*.
Markov Decision Processes in Artificial Intelligence

SOLNON Christine
Ant Colony Optimization and Constraint Programming

AUBRUN Christophe, SIMON Daniel, SONG Ye-Qiong *et al.*
Co-design Approaches for Dependable Networked Control Systems

2009

FOURNIER Jean-Claude
Graph Theory and Applications

GUEDON Jeanpierre
The Mojette Transform / Theory and Applications

JARD Claude, ROUX Olivier
Communicating Embedded Systems / Software and Design

LECOUTRE Christophe
Constraint Networks / Targeting Simplicity for Techniques and Algorithms

2008

BANÂTRE Michel, MARRÓN Pedro José, OLLERO Hannibal, WOLITZ Adam
Cooperating Embedded Systems and Wireless Sensor Networks

MERZ Stephan, NAVET Nicolas
Modeling and Verification of Real-time Systems

PASCHOS Vangelis Th
Combinatorial Optimization and Theoretical Computer Science: Interfaces and Perspectives

WALDNER Jean-Baptiste
Nanocomputers and Swarm Intelligence

2007

BENHAMOU Frédéric, JUSSIEN Narendra, O'SULLIVAN Barry
Trends in Constraint Programming

JUSSIEN Narendra
A TO Z OF SUDOKU

2006

BABAU Jean-Philippe *et al.*
From MDD Concepts to Experiments and Illustrations – DRES 2006

HABRIAS Henri, FRAPPIER Marc
Software Specification Methods

MURAT Cecile, PASCHOS Vangelis Th
Probabilistic Combinatorial Optimization on Graphs

PANETTO Hervé, BOUDJLIDA Nacer
Interoperability for Enterprise Software and Applications 2006 / IFAC-IFIP I-ESA'2006

2005

GÉRARD Sébastien *et al.*
Model Driven Engineering for Distributed Real Time Embedded Systems

PANETTO Hervé
Interoperability of Enterprise Software and Applications 2005

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.

Big data analytics is one of the fastest growing fields in today's business context. It concerns not only data that is continuously getting "bigger", but also the ways in which that data is processed and analyzed and how it is turned into knowledge.

This book covers the basic concepts of big data analytics in easy-to-understand terms, working through a series of practical applications which illustrate the use and importance of working with data. It includes an introduction to the crucial domain of machine learning algorithms based on their principles, concepts and applications.

As companies are increasingly embracing the opportunities offered by big data, it is becoming ever more important to understand what exactly big data analytics and algorithms can do. This book will help the reader to explore this pertinent new field and discover what can be revealed by each byte of data.

Soraya Sedkaoui is a Senior Lecturer at Khemis Miliana University, Algeria, and a Data Analyst with SRY Consulting, France. Her research fields include econometrics, big data, computer science and the development of algorithms and models for business applications.

www.iste.co.uk

WILEY

9 781786 303264