1

What Can We Do With Data?

Until recently, researchers working with data analysis were struggling to obtain data for their experiments. Recent advances in the technology of data processing, data storage and data transmission, associated with advanced and intelligent computer software, reducing costs and increasing capacity, have changed this scenario. It is the time of the Internet of Things, where the aim is to have everything or almost everything connected. Data previously produced on paper are now on-line. Each day, a larger quantity of data is generated and consumed. Whenever you place a comment in your social network, upload a photograph, some music or a video, navigate through the Internet, or add a comment to an e-commerce web site, you are contributing to the data increase. Additionally, machines, financial transactions and sensors such as security cameras, are increasingly gathering data from very diverse and widespread sources.

In 2012, it was estimated that, each year, the amount of data available in the world doubles [1]. Another estimate, from 2014, predicted that by 2020 all information will be digitized, eliminated or reinvented in 80% of processes and products of the previous decade [2]. In a third report, from 2015, it was predicted that mobile data traffic will be almost 10 times larger in 2020 [3]. The result of all these rapid increases of data is named by some the "data explosion".

Despite the impression that this can give – that we are drowning in data – there are several benefits from having access to all these data. These data provide a rich source of information that can be transformed into new, useful, valid and human-understandable knowledge. Thus, there is a growing interest in exploring these data to extract this knowledge, using it to support decision making in a wide variety of fields: agriculture, commerce, education, environment, finance, government, industry, medicine, transport and social care. Several companies around the world are realizing the gold mine they have and the potential of these data to support their work, reduce waste and dangerous and tedious work activities, and increase the value of their products and their profits.

A General Introduction to Data Analytics, First Edition. João Mendes Moreira, André C. P. L. F. de Carvalho, and Tomáš Horváth.

Companion website: www.wiley.com/go/moreira/dataanalytics

The analysis of these data to extract such knowledge is the subject of a vibrant area known as data analytics, or simply "analytics". You can find several definitions of analytics in the literature. The definition adopted here is:

Analytics The science that analyze crude data to extract useful knowledge (patterns) from them.

This process can also include data collection, organization, pre-processing, transformation, modeling and interpretation.

Analytics as a knowledge area involves input from many different areas. The idea of generalizing knowledge from a data sample comes from a branch of statistics known as inductive learning, an area of research with a long history. With the advances of personal computers, the use of computational resources to solve problems of inductive learning become more and more popular. Computational capacity has been used to develop new methods. At the same time, new problems have appeared requiring a good knowledge of computer sciences. For instance, the ability to perform a given task with more computational efficiency has become a subject of study for people working in computational statistics.

In parallel, several researchers have dreamed of being able to reproduce human behavior using computers. These were people from the area of artificial intelligence. They also used statistics for their research but the idea of reproducing human and biological behavior in computers was an important source of motivation. For instance, reproducing how the human brain works with artificial neural networks has been studied since the 1940s; reproducing how ants work with ant colony optimization algorithm since the 1990s. The term machine learning (ML) appeared in this context as the "field of study that gives computers the ability to learn without being explicitly programmed," according to Arthur Samuel in 1959 [4].

In the 1990s, a new term appeared with a different slight meaning: data mining (DM). The 1990s was the decade of the appearance of business intelligence tools as consequence of the data facilities having larger and cheaper capacity. Companies start to collect more and more data, aiming to either solve or improve business operations, for example by detecting frauds with credit cards, by advising the public of road network constraints in cities, or by improving relations with clients using more efficient techniques of relational marketing. The question was of being able to mine the data in order to extract the knowledge necessary for a given task. This is the goal of data mining.

1.1 Big Data and Data Science

In the first years of the 20th century, the term big data has appeared. Big data, a technology for data processing, was initially defined by the "three Vs", although some more Vs have been proposed since. The first three Vs allow us to define

a taxonomy of big data. They are: volume, variety and velocity. Volume is concerned with how to store big data: data repositories for large amounts of data. Variety is concerned with how to put together data from different sources. Velocity concerns the ability to deal with data arriving very fast, in streams known as data streams. Analytics is also about discovering knowledge from data streams, going beyond the velocity component of big data.

Another term that has appeared and is sometimes used as a synonym for big data is data science. According to Provost and Fawcett [5], big data are data sets that are too large to be managed by conventional data-processing technologies, requiring the development of new techniques and tools for data storage, processing and transmission. These tools include, for example, MapReduce, Hadoop, Spark and Storm. But data volume is not the only characterization of big data. The word "big" can refer to the number of data sources, to the importance of the data, to the need for new processing techniques, to how fast data arrive, to the combination of different sets of data so they can be analyzed in real time, and its ubiquity, since any company, nonprofit organization or individual has access to data now.

Thus big data is more concerned with technology. It provides a computing environment, not only for analytics, but also for other data processing tasks. These tasks include finance transaction processing, web data processing and georeferenced data processing.

Data science is concerned with the creation of models able to extract patterns from complex data and the use of these models in real-life problems. Data science extracts meaningful and useful knowledge from data, with the support of suitable technologies. It has a close relationship to analytics and data mining. Data science goes beyond data mining by providing a knowledge extraction framework, including statistics and visualization.

Therefore, while big data gives support to data collection and management, data science applies techniques to these data to discover new and useful knowledge: big data collects and data science discovers. Other terms such as knowledge discovery or extraction, pattern recognition, data analysis, data engineering, and several others are also used. The definition we use of data analytics covers all these areas that are used to extract knowledge from data.

1.2 Big Data Architectures

As data increase in size, velocity and variety, new computer technologies become necessary. These new technologies, which include hardware and software, must be easily expanded as more data are processed. This property is known as scalability. One way to obtain scalability is by distributing the data processing tasks into several computers, which can be combined into clusters of computers. The reader should not confuse clusters of computers

with clusters produced by clustering techniques, which are techniques from analytics in which a data set is partitioned to find groups within it.

Even if processing power is expanded by combining several computers in a cluster, creating a distributed system, conventional software for distributed systems usually cannot cope with big data. One of the limitations is the efficient distribution of data among the different processing and storage units. To deal with these requirements, new software tools and techniques have been developed.

One of the first techniques developed for big data processing using clusters was MapReduce. MapReduce is a programming model that has two steps: map and reduce. The most famous implementation of MapReduce is called Hadoop.

MapReduce divides the data set into parts – chunks – and stores in the memory of each cluster computer the chunk of the data set needed by this computer to accomplish its processing task. As an example, suppose that you need to calculate the average salary of 1 billion people and you have a cluster with 1000 computers, each with a processing unit and a storage memory. The people can be divided into 1000 chunks – subsets – with data from 1 million people each. Each chunk can be processed independently by one of the computers. The results produced by each these computers (the average salary of 1 million people) can be averaged, returning the final salary average.

To efficiently solve a big data problem, a distributed system must attend the following requirements:

- Make sure that no chunk of data is lost and the whole task is concluded. If one or more computers has a failure, their tasks, and the corresponding data chunk, must be assumed by another computer in the cluster.
- Repeat the same task, and corresponding data chunk, in more than one cluster computer; this is called redundancy. Thus, if one or more computer fails, the redundant computer carries on with the task.
- Computers that have had faults can return to the cluster again when they are fixed.
- Computers can be easily removed from the cluster or extra ones included in it as the processing demand changes.

A solution incorporating these requirements must hide from the data analyst the details of how the software works, such as how the data chunks and tasks are divided among the cluster computers.

1.3 Small Data

In the opposite direction from big data technologies and methods, there is a movement towards more personal, subjective analysis of chunks of data, termed "small data". Small data is a data set whose volume and format allows its processing and analysis by a person or a small organization. Thus, instead of

collecting data from several sources, with different formats, and generated at increasing velocities, creating large data repositories and processing facilities, small data favors the partition of a problem into small packages, which can be analyzed by different people or small groups in a distributed and integrated way.

People are continuously producing small data as they perform their daily activities, be it navigating the web, buying a product in a shop, undergoing medical examinations and using apps in their mobiles. When these data are collected to be stored and processed in large data servers they become big data. To be characterized as small data, a data set must have a size that allows its full understanding by an user.

The type of knowledge sought in big and small data is also different, with the first looking for correlations and the second for causality relations. While big data provide tools that allow companies to understand their customers, small data tools try to help customers to understand themselves. Thus, big data is concerned with customers, products and services, and small data is concerned with the individuals that produced the data.

1.4 What is Data?

But what is data about? Data, in the information age, are a large set of bits encoding numbers, texts, images, sounds, videos, and so on. Unless we add information to data, they are meaningless. When we add information, giving a meaning to them, these data become knowledge. But before data become knowledge, typically, they pass through several steps where they are still referred to as data, despite being a bit more organized; that is, they have *some* information associated with them.

Let us see the example of data collected from a private list of acquaintances or contacts.

Information as presented in Table 1.1, usually referred to as tabular data, is characterized by the way data are organized. In tabular data, data are organized in rows and columns, where each column represents a characteristic of the data and each row represents an occurrence of the data. A column is referred to as an attribute or, with the same meaning, a feature, while a row is referred to as an instance, or with the same meaning, an object.

Instance or Object Examples of the concept we want to characterize.

Example 1.1 In the example in Table 1.1, we intend to characterize people in our private contact list. Each member is, in this case, an instance or object. It corresponds to a row of the table.

Attribute or Feature Attributes, also called features, are characteristics of the instances.

Table 1.1 Data set of our private contact list.

Contact	Age	Educational level	Company
Andrew	55	1.0	Good
Bernhard	43	2.0	Good
Carolina	37	5.0	Bad
Dennis	82	3.0	Good
Eve	23	3.2	Bad
Fred	46	5.0	Good
Gwyneth	38	4.2	Bad
Hayden	50	4.0	Bad
Irene	29	4.5	Bad
James	42	4.1	Good
Kevin	35	4.5	Bad
Lea	38	2.5	Good
Marcus	31	4.8	Bad
Nigel	71	2.3	Good

Example 1.2 In Table 1.1, contact, age, education level and company are four different attributes.

The majority of the chapters in this book expect the data to be in tabular format; that is, already organized by rows and columns, each row representing an instance and each column representing an attribute. However, a table can be organized differently, having the instances per column and the attributes per row.

There are, however, data that are not possible to represent in a single table.

Example 1.3 As an example, if some of the contacts are relatives of other contacts, a second table, as shown in Table 1.2, representing the family relationships, would be necessary. You should note that each person referred to in Table 1.2 also exists in Table 1.1, i.e., there are relations between attributes of different tables.

Data sets represented by several tables, making clear the relations between these tables, are called relational data sets. This information is easily handled using relational databases. In this book, only simple forms of relational data will be used. This is discussed in each chapter whenever necessary.

Table 1.2 Family relations between contacts.

Friend	Father	Mother	Sister
Eve	Andrew	Hayden	Irene
Irene	Andrew	Hayden	Eve

Example 1.4 In our example, data is split into two tables, one with the individual data of each contact (Table 1.1) and the other with the data about the family relations between them (Table 1.2).

1.5 A Short Taxonomy of Data Analytics

Now that we know what data are, we will look at what we can do with them. A natural taxonomy that exists in data analytics is:

- Descriptive analytics: summarize or condense data to extract patterns
- Predictive analytics: extract models from data to be used for future predictions.

In descriptive analytics tasks, the result of a given method or technique,[1] is obtained directly by applying an algorithm to the data. The result can be a statistic, such as an average, a plot, or a set of groups with similar instances, among other things, as we will see in this book. Let us see the definition of method and algorithm.

Method or technique A method or technique is a systematic procedure that allows us to achieve an intended goal.

A method shows how to perform a given task. But in order to use a language closer to the language computers can understand, it is necessary to describe the method/technique through an algorithm.

Algorithm An algorithm is a self-contained, step-by-step set of instructions easily understandable by humans, allowing the implementation of a given method. They are self-contained in order to be easily translated to an arbitrary programming language.

Example 1.5 The method to obtain the average age of my contacts uses the ages of each (we could use other methods, such as using the number of contacts for each different age). A possible algorithm for this very simple example is shown next.

1 These two terms are used interchangeably in this book.

Algorithm An algorithm to calculate the average age of our contacts

1: INPUT: A: a vector of size N with the ages of all contacts.
2: $S \leftarrow 0$ ▷ Initialize the sum S to zero
3: **for** $i = 1$ **to** N **do** ▷ Iterate through all the elements of A.
4: $S \leftarrow S + A_i$ ▷ Add the current (ith) element of A to S.
5: $\overline{A} \leftarrow S/N$ ▷ Divide the sum by the number N of contacts.
6: return($\overline{A}$) ▷ Return the result, i.e. the average age of the N contacts.

In the limit, a method can be straightforward. It is possible, in many cases, to express it as a formula instead of as an algorithm.

Example 1.6 For instance, the average could be expressed as: $\overline{A} = \sum_{i=1}^{N} A_i/N$.

We have seen an algorithm that describes a descriptive method. An algorithm can also describe predictive methods. In this last case it describes how to generate a model. Let us see what a model is.

Model A model in data analytics is a generalization obtained from data that can be used afterwords to generate predictions for new given instances. It can be seen as a prototype that can be used to make predictions. Thus, model induction is a predictive task.

Example 1.7 If we apply an algorithm for induction of decision trees to provide an explanation of who, among our contacts, is a good company, we obtain a model, called a decision tree, like the one presented in Figure 1.1. It can be seen that people older than 38 years are typically better company than those whose age is equal or less than 38 more than 80% of people aged 38 or less are bad company, while more than 80% of people older than 38 are good company. This model could be used to predict whether a new contact is or not a good company. It would be enough to know the age of that new contact.

Now that we have a rough idea of what analytics is, let us see real examples of problems in data analytics.

1.6 Examples of Data Use

We will describe two real-world problems from different areas as an introduction to the different subjects that are covered in this book. Many more could be presented. One of the problems is from medicine and the other is

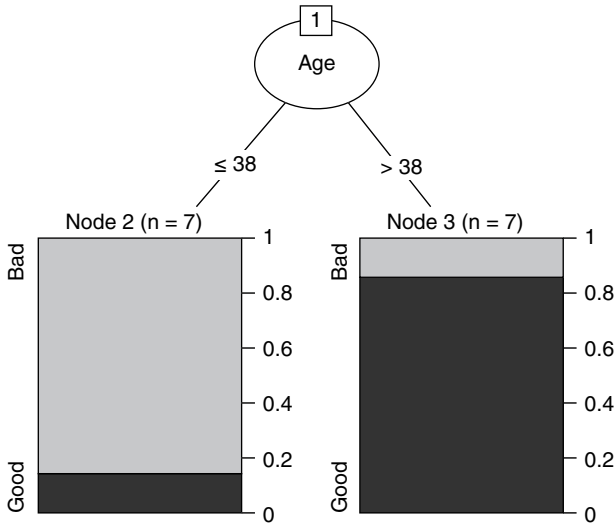

Figure 1.1 A prediction model to classify someone as either good or bad company.

from economics. The problems were chosen with a view to the availability of relevant data, because the problems involved will be solved in the project chapters of the book (Chapters 7 and 12).

1.6.1 Breast Cancer in Wisconsin

Breast cancer is a well-known problem that affects mainly women. The detection of breast tumors can be performed through a biopsy technique known as fine-needle aspiration. This uses a fine needle to sample cells from the mass under study. Samples of breast mass obtained using fine-needle aspiration were recorded in a set of images [6]. Then, a dataset was collected by extracting features from these images. The objective of the first problem is to detect different patterns of breast tumors in this dataset, to enable it to be used for diagnostic purposes.

1.6.2 Polish Company Insolvency Data

The second problem concerns the prediction of the economic wealth of Polish companies. Can we predict which companies will become insolvent in the next five years? The answer to this question is obviously relevant to institutions and shareholders.

1.7 A Project on Data Analytics

Every project needs a plan. Or, to be precise, a methodology to prepare the plan. A project on data analytics does not imply only the use of one or more specific methods. It implies:

- understanding the problem to be solved
- defining the objectives of the project
- looking for the necessary data
- preparing these data so that they can be used
- identifying suitable methods and choosing between them
- tuning the hyper-parameters of each method (see below)
- analyzing and evaluating the results
- redoing the pre-processing tasks and repeating the experiments
- and so on.

In this book, we assume that in the induction of a model, there are both hyper-parameters and parameters whose values are set. The values of the hyper-parameters are set by the user, or some external optimization method. The parameter values, on the other hand, are model parameters whose values are set by a modeling or learning algorithm in its internal procedure. When the distinction is not clear, we use the term parameter. Thus, hyper-parameters might be, for example, the number of layers and the activation function in a multi-layer perceptron neural network and the number of clusters for the k-means algorithm. Examples of parameters are the weights found by the backpropagation algorithm when training a multi-layer perceptron neural network and the distribution of objects carried out by k-means. Multi-layer perceptron neural networks and k-means will be explained later in this book.

How can we perform all these operations in an organized way? This section is all about methodologies for planning and developing projects in data analytics.

A brief history of methodologies for data analytics is presented first. Afterwards, two different methodologies are described:

- a methodology from Academia, KDD
- a methodology from industry, CRISP-DM.

The latter is used in the cheat sheet and project chapters (Chapters 7 and 12).

1.7.1 A Little History on Methodologies for Data Analytics

Machine learning, knowledge discovery from data and related areas experienced strong development in the 1990s. Both in academia and industry, the research on these topics was advancing quickly. Naturally, methodologies for projects in these areas, now referred to as data analytics, become a necessity.

In the mid-1990s, both in academia and industry, different methodologies were presented.

The most successful methodology from academia came from the USA. This was the KDD process of Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth [7]. Despite being from academia, the authors had considerable work experience in industry.

The most successful tool from industry, was and still is the CRoss-Industry Standard Process for Data Mining (CRISP-DM) [8]. Conceived in 1996, it later got underway as an European Union project under the ESPRIT funding initiative. The project had five partners from industry: SPSS, Teradata, Daimler AG, NCR Corporation and OHRA, an insurance company. In 1999 the first version was presented. An attempt to create a new version began between 2006 and 2008 but no new discoveries are known from these efforts. CRISP-DM is nowadays used by many different practitioners and by several corporations, in particular IBM. However, despite its popularity, CRISP-DM needs new developments in order to meet the new challenges from the age of big data.

Other methodologies exist. Some of them are domain-specific: they assume the use of a given tool for data analytics. This is not the case for SEMMA, which, despite has been created by SAS, is tool independent. Each letter of its name, SEMMA, refers to one of its five steps: Sample, Explore, Modify, Model and Assess.

Polls performed by kdnuggets [9] over the years (2002, 2004, 2007 and 2014) show how methodologies on data analytics have been used through time (Figure 1.2).

Next, the KDD process and the CRISP-DM methodologies are described in detail.

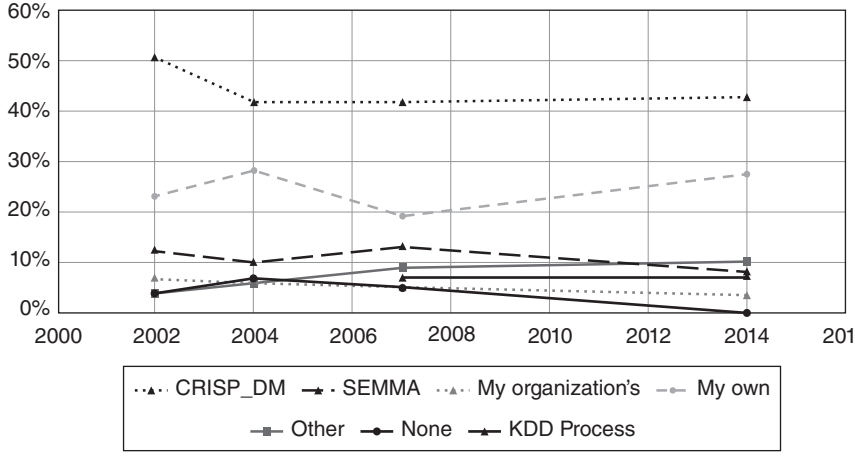

Figure 1.2 The use of different methodologies on data analytics through time.

1.7.2 The KDD Process

Intended to be a methodology that could cope with all the processes necessary to extract knowledge from data, the KDD process proposes a sequence of nine steps. In spite of the sequence, the KDD process considers the possibility of going back to any previous step in order to redo some part of the process. The nine steps are:

1) *Learning the application domain:* What is expected in terms of the application domain? What are the characteristics of the problem; its specificities? A good understanding of the application domain is required.
2) *Creating a target dataset:* What data are needed for the problem? Which attributes? How will they be collected and put in the desired format (say, a tabular data set)? Once the application domain is known, the data analyst team should be able to identify the data necessary to accomplish the project.
3) *Data cleaning and pre-processing:* How should missing values and/or outliers such as extreme values be handled? What data type should we choose for each attribute? It is necessary to put the data in a specific format, such as a tabular format.
4) *Data reduction and projection:* Which features should we include to represent the data? From the available features, which ones should be discarded? Should further information be added, such as adding the day of the week to a timestamp? This can be useful in some tasks. Irrelevant attributes should be removed.
5) *Choosing the data mining function:* Which type of methods should be used? Four types of method are: summarization, clustering, classification and regression. The first two are from the branch of descriptive analytics while the latter two are from predictive analytics.
6) *Choosing the data mining algorithm(s):* Given the characteristics of the problem and the characteristics of the data, which methods should be used? It is expected that specific algorithms will be selected.
7) *Data mining:* Given the characteristics of the problem, the characteristics of the data, and the applicable method type, which specific methods should be used? Which values should be assigned to the hyper-parameters? The choice of method depends on many different factors: interpretability, ability to handle missing values, capacity to deal with outliers, computational efficiency, among others.
8) *Interpretation:* What is the meaning of the results? What is the utility for the final user? To select the useful results and to evaluate them in terms of the application domain is the goal of this step. It is common to go back to a previous step when the results are not as good as expected.
9) *Using discovered knowledge:* How can we apply the new knowledge in practice? How is it integrated in everyday life? This implies the integration of the new knowledge into the operational system or in the reporting system.

For simplicity sake, the nine steps were described sequentially, which is typical. However, in practice, some jumps are often necessary. As an example, steps 3 and 4 can be grouped together with steps 5 and 6. The way we pre-process the data depends on the methods we will use. For instance, some methods are able to deal with missing values, others not. When a method is not able to deal with missing values, those missing values should be included somehow or some attributes or instances should be removed. Also, there are methods that are too sensitive to outliers or extreme values. When this happens, outliers should be removed. Otherwise, it is not necessary to remove them. These are just examples on how data cleaning and pre-processing tasks depend on the chosen method(s) (steps 5 and 6).

1.7.3 The CRISP-DM Methodology

CRoss-Industry Standard Process for Data Mining (CRISP-DM) is a six-step method, which, like the KDD process, uses a non-rigid sequential framework. Despite the six phases, CRISP-DM is seen as a perpetual process, used throughout the life of a company in successive iterations (Figure 1.3).

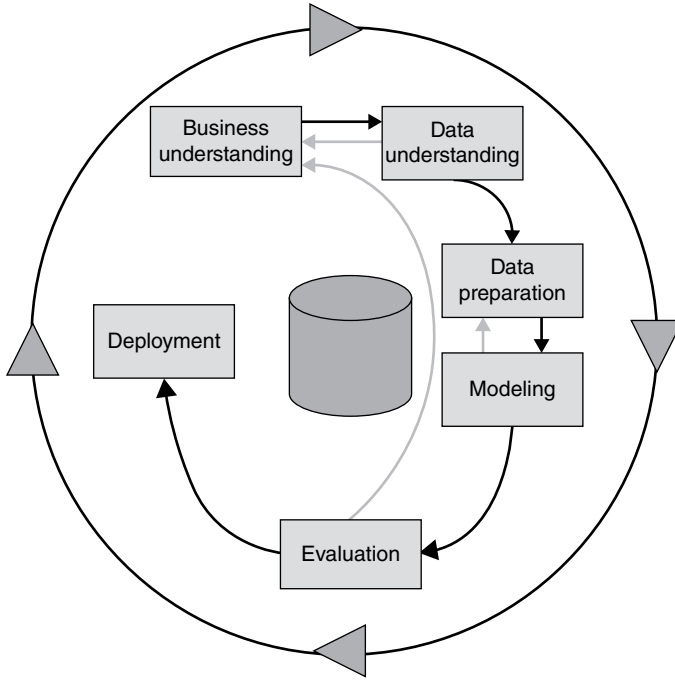

Figure 1.3 The CRISP-DM methodology (adapted from http://www.crisp-dm.org/).

The six phases are:

1) *Business understanding:* This involves understanding the business domain, being able to define the problem from the business domain perspective, and finally being able to translate such business problems into a data analytics problem.
2) *Data understanding:* This involves collection of the necessary data and their initial visualization/summarization in order to obtain the first insights, particularly but not exclusively, about data quality problems such as missing data or outliers.
3) *Data preparation:* This involves preparing the data set for the modeling tool, and includes data transformation, feature construction, outlier removal, missing data fulfillment and incomplete instances removal.
4) *Modeling:* Typically there are several methods that can be used to solve the same problem in analytics, often with specific data requirements. This implies that there may be a need for additional data preparation tasks that are method specific. In such case it is necessary to go back to the previous step. The modeling phase also includes tuning the hyper-parameters for each of the chosen method(s).
5) *Evaluation:* Solving the problem from the data analytics point of view is not the end of the process. It is now necessary to understand how its use is meaningful from the business perspective; in other words, that the obtained solution answers to the business requirements.
6) *Deployment:* The integration of the data analytics solution in the business process is the main purpose of this phase. Typically, it implies the integration of the obtained solution into a decision-support tool, website maintenance process, reporting process or elsewhere.

A more comprehensive description of the CRISP-DM methodology is presented in Appendix A. This will certainly be useful to help you to develop the projects at the end of each of Parts II and III of the book, as explained in Section 1.8.

1.8 How this Book is Organized

The book has two main parts: on descriptive (Part II) and predictive (Part III) analytics respectively.

Parts II and III will finish with a cheat sheet and project chapter (Chapter 7 for Part II, and Chapter 12 for Part III) where the contents of each part are summarized and a project is proposed using one of the two real-world problems presented above (Section 1.6). These projects will be developed using the CRISP-DM methodology, as described in Section 1.7.3 and Appendix A, the latter being a more detailed description. In all other chapters, including this one, we will use as example a small data set from an idealized private list of

contacts in order to better explain the methods. The data set will be presented in the chapters as necessary.

All chapters, excluding this one, the cheat sheet and project chapters, will have exercises. In this book there is no specific software for the examples and exercises. This book was conceived of as a 13-week course, covering one chapter per week. The content of each part is described next.

Part I includes the present chapter, where introductory concepts, a brief methodological description and some examples are presented.

Part II presents the main methods of descriptive analytics and data preprocessing. There are five families of methods/tools covered; one per chapter. The first one, in Chapter 2, is on descriptive statistics. It aims to describe data in a way that us humans can better extract knowledge from. However, the methods described only apply to data with a maximum of two attributes. Chapter 3 extends the discussion in Chapter 2 to an arbitrary number of attributes. The methods described are known as multivariate descriptive statistics methods. Chapter 4 describes methods that are typically used in the data preparation phase of the CRISP-DM methodology, concerning data quality issues, converting data to different scales or scale types and reducing data dimensionality. Chapter 5 describes methods involving clustering, an important technique that aims to find groups of similar instances. Clustering is used in a large number of fields, most notably marketing, in order to obtain segments of clients with similar behavior. Chapter 6 is about a family of descriptive methods known as frequent pattern mining, which aims to capture the most frequent patterns. It is particularly common in the retail market, where it is used in market basket analysis.

Part III presents the main methods of predictive analytics. Chapter 8 is about regression; that is, the prediction of a quantitative attribute. It covers generalization, performance measures for regression and the bias–variance trade-off. It also presents some of the most popular algorithms for regression: multivariate linear regression, ridge and lasso regression, principal component regression and partial least squares regression. In Chapter 9, the binary classification problem is introduced, together with performance measures for classification and methods based on probabilities and distance measures. In Chapter 10 more advanced and state-of-the-art methods of prediction are described: decision trees, artificial neural networks and support vector machines. Chapter 11 presents the most popular algorithms for ensemble learning. Then, a discussion on algorithm bias is presented. Classification tasks other than than binary classification are discussed, as well as other topics relevant for prediction such as imbalanced data classification, semi-supervised learning and active learning. Finally, a discussion on the use of supervised interpretable techniques for descriptive and predictive analysis is presented.

Part IV has only Chapter 13, which discusses briefly applications for text, the web and social media, using both descriptive and predictive approaches.

1.9 Who Should Read this Book

Anyone who aims to extract knowledge from data, whatever they are, could and should read this book. This is a book where the main concepts of data analytics can be understood, and not only by people with a background in engineering and the exact sciences.

You do not need to know statistics – but it helps – or programming. You do not need to be a student of computer sciences, or even a student! You only need to study a little. This book was conceived of as being for bachelor's or master's students, levels where these kinds of analytic tools are relevant. In our experience, more and more people are interested in analyzing data. So this book was written in order to introduce the main tools for data analytics. It aims to be understandable by any university student whichever their background is.

It is expected that after reading this book you will be able to develop a project on analytics, assuming that you are already familiar with the business area of the project. You should be able to identify the necessary data, pre-process and clean it, choose the methods suitable to the project, tune and apply them, evaluate the results in terms of the project purpose, and give the necessary instructions to a development team for deployment.

In order to suit an audience without a background in computer science and/or quantitative methods, the book is particularly focused on explaining the concepts in an intuitive way. Whenever possible, graphics are used to explain the methods. Special attention is given to what must be considered in order to make the right methodological choices. For instance, knowing the meaning of a hyper-parameter allow us to define a strategy for tuning its value.

In summary, it is not expected that after reading this book you will be able to develop new methods or algorithms. But is it expected that you can correctly use appropriate methods to deal with data analytics problems.