Introduction to Data Science

1

1.1 What is Data Science?

You have, no doubt, already experienced data science in several forms. When you are looking for information on the web by using a search engine or asking your mobile phone for directions, you are interacting with data science products. Data science has been behind resolving some of our most common daily tasks for several years.

Most of the scientific methods that power data science are not new and they have been out there, waiting for applications to be developed, for a long time. Statistics is an old science that stands on the shoulders of eighteenth-century giants such as Pierre Simon Laplace (1749–1827) and Thomas Bayes (1701–1761). Machine learning is younger, but it has already moved beyond its infancy and can be considered a well-established discipline. Computer science changed our lives several decades ago and continues to do so; but it cannot be considered new.

So, why is data science seen as a novel trend within business reviews, in technology blogs, and at academic conferences?

The novelty of data science is not rooted in the latest scientific knowledge, but in a disruptive change in our society that has been caused by the evolution of technology: datification. Datification is the process of rendering into data aspects of the world that have never been quantified before. At the personal level, the list of datified concepts is very long and still growing: business networks, the lists of books we are reading, the films we enjoy, the food we eat, our physical activity, our purchases, our driving behavior, and so on. Even our thoughts are datified when we publish them on our favorite social network; and in a not so distant future, your gaze could be datified by wearable vision registering devices. At the business level, companies are datifying semi-structured data that were previously discarded: web activity logs, computer network activity, machinery signals, etc. Nonstructured data, such as written reports, e-mails, or voice recordings, are now being stored not only for archive purposes but also to be analyzed.

L. Igual and S. Seguí, *Introduction to Data Science*,
Undergraduate Topics in Computer Science, DOI 10.1007/978-3-319-50017-1_1

However, datification is not the only ingredient of the data science revolution. The other ingredient is the democratization of data analysis. Large companies such as Google, Yahoo, IBM, or SAS were the only players in this field when data science had no name. At the beginning of the century, the huge computational resources of those companies allowed them to take advantage of datification by using analytical techniques to develop innovative products and even to take decisions about their own business. Today, the analytical gap between those companies and the rest of the world (companies and people) is shrinking. Access to cloud computing allows any individual to analyze huge amounts of data in short periods of time. Analytical knowledge is free and most of the crucial algorithms that are needed to create a solution can be found, because open-source development is the norm in this field. As a result, the possibility of using rich data to take evidence-based decisions is open to virtually any person or company.

Data science is commonly defined as a methodology by which actionable insights can be inferred from data. This is a subtle but important difference with respect to previous approaches to data analysis, such as business intelligence or exploratory statistics. Performing data science is a task with an ambitious objective: the production of beliefs informed by data and to be used as the basis of decision-making. In the absence of data, beliefs are uninformed and decisions, in the best of cases, are based on best practices or intuition. The representation of complex environments by rich data opens up the possibility of applying all the scientific knowledge we have regarding how to infer knowledge from data.

In general, data science allows us to adopt four different strategies to explore the world using data:

1. *Probing reality*. Data can be gathered by passive or by active methods. In the latter case, data represents the response of the world to our actions. Analysis of those responses can be extremely valuable when it comes to taking decisions about our subsequent actions. One of the best examples of this strategy is the use of A/B testing for web development: What is the best button size and color? The best answer can only be found by probing the world.
2. *Pattern discovery*. Divide and conquer is an old heuristic used to solve complex problems; but it is not always easy to decide how to apply this common sense to problems. Datified problems can be analyzed automatically to discover useful patterns and natural clusters that can greatly simplify their solutions. The use of this technique to profile users is a critical ingredient today in such important fields as programmatic advertising or digital marketing.
3. *Predicting future events*. Since the early days of statistics, one of the most important scientific questions has been how to build robust data models that are capable of predicting future data samples. Predictive analytics allows decisions to be taken in response to future events, not only reactively. Of course, it is not possible to predict the future in any environment and there will always be unpredictable events; but the identification of predictable events represents valuable knowledge. For example, predictive analytics can be used to optimize the tasks

planned for retail store staff during the following week, by analyzing data such as weather, historic sales, traffic conditions, etc.

4. *Understanding people and the world*. This is an objective that at the moment is beyond the scope of most companies and people, but large companies and governments are investing considerable amounts of money in research areas such as understanding natural language, computer vision, psychology and neuroscience. Scientific understanding of these areas is important for data science because in the end, in order to take optimal decisions, it is necessary to know the real processes that drive people's decisions and behavior. The development of deep learning methods for natural language understanding and for visual object recognition is a good example of this kind of research.

1.2 About This Book

Data science is definitely a cool and trendy discipline that routinely appears in the headlines of very important newspapers and on TV stations. Data scientists are presented in those forums as a scarce and expensive resource. As a result of this situation, data science can be perceived as a complex and scary discipline that is only accessible to a reduced set of geniuses working for major companies. The main purpose of this book is to demystify data science by describing a set of tools and techniques that allows a person with basic skills in computer science, mathematics, and statistics to perform the tasks commonly associated with data science.

To this end, this book has been written under the following assumptions:

- Data science is a complex, multifaceted field that can be approached from several points of view: ethics, methodology, business models, how to deal with big data, data engineering, data governance, etc. Each point of view deserves a long and interesting discussion, but the approach adopted in this book focuses on analytical techniques, because such techniques constitute the core toolbox of every data scientist and because they are the key ingredient in predicting future events, discovering useful patterns, and probing the world.
- You have some experience with Python programming. For this reason, we do not offer an introduction to the language. But even if you are new to Python, this should not be a problem. Before reading this book you should start with any online Python course. Mastering Python is not easy, but acquiring the basics is a manageable task for anyone in a short period of time.
- Data science is about evidence-based storytelling and this kind of process requires appropriate tools. The Python data science toolbox is one, not the only, of the most developed environments for doing data science. You can easily install all you need by using Anaconda[1]: a free product that includes a programming language

[1] https://www.continuum.io/downloads.

(Python), an interactive environment to develop and present data science projects (Jupyter notebooks), and most of the toolboxes necessary to perform data analysis.
- Learning by doing is the best approach to learn data science. For this reason all the code examples and data in this book are available to download at https://github.com/DataScienceUB/introduction-datascience-python-book.
- Data science deals with solving real-world problems. So all the chapters in the book include and discuss practical cases using real data.

This book includes three different kinds of chapters. The first kind is about Python extensions. Python was originally designed to have a minimum number of data objects (int, float, string, etc.); but when dealing with data, it is necessary to extend the native set to more complex objects such as (numpy) numerical arrays or (pandas) data frames. The second kind of chapter includes techniques and modules to perform statistical analysis and machine learning. Finally, there are some chapters that describe several applications of data science, such as building recommenders or sentiment analysis. The composition of these chapters was chosen to offer a panoramic view of the data science field, but we encourage the reader to delve deeper into these topics and to explore those topics that have not been covered: big data analytics, deep learning techniques, and more advanced mathematical and statistical methods (e.g., computational algebra and Bayesian statistics).

Acknowledgements This chapter was co-written by Jordi Vitrià.