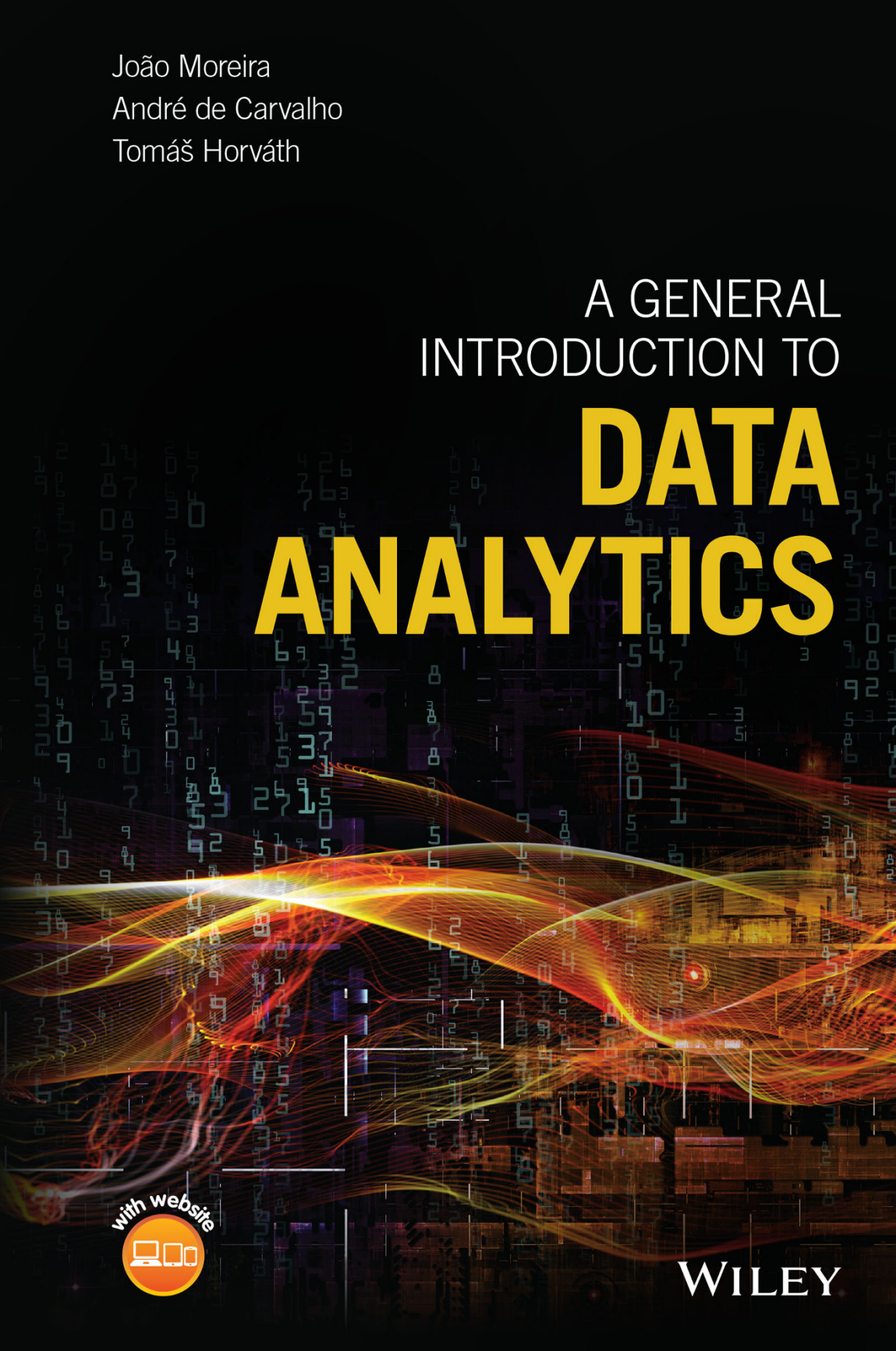
João Moreira
André de Carvalho
Tomáš Horváth

# A GENERAL INTRODUCTION TO DATA ANALYTICS




WILEY

# A General Introduction to Data Analytics

# A General Introduction to Data Analytics

*João Mendes Moreira*
*University of Porto*

*André C. P. L. F. de Carvalho*
*University of São Paulo*

*Tomáš Horváth*
*Eötvös Loránd University in Budapest*
*Pavol Jozef Šafárik University in Košice*

WILEY

This edition first published 2019


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

*Editorial Office*
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.



*Library of Congress Cataloging-in-Publication Data*

Names: Moreira, João, 1969– author. | Carvalho, André Carlos Ponce de Leon Ferreira, author. | Horváth, Tomáš, 1976– author.
Title: A general introduction to data analytics / by João Mendes Moreira, André C. P. L. F. de Carvalho, Tomáš Horváth.
Description: Hoboken, NJ : John Wiley & Sons, 2019. | Includes bibliographical references and index. |
Identifiers: LCCN 2017060728 (print) | LCCN 2018005929 (ebook) | ISBN 9781119296256 (pdf) | ISBN 9781119296263 (epub) | ISBN 9781119296249 (cloth)
Subjects: LCSH: Mathematical statistics–Methodology. | Electronic data processing. | Data mining.
Classification: LCC QA276.4 (ebook) | LCC QA276.4 .M664 2018 (print) | DDC 519.50285–dc23
LC record available at https://lccn.loc.gov/2017060728

Cover image: © agsandrew/Shutterstock
Cover design by Wiley

Printed in the United States of America.

Set in 10/12pt Warnock by SPi Global, Pondicherry, India

10 9 8 7 6 5 4 3 2 1

*To the women at home that make my life better: Mamã, Yá and Yé – João*
*To my family, Valeria, Beatriz, Gabriela and Mariana – André*
*To my wife Danielle – Tomáš*

# Contents

# Preface

We are living in a period of history that will certainly be remembered as one where information began to be instantaneously obtainable, services were tailored to individual criteria, and people did what made them feel good (if it did not put their lives at risk). Every year, machines are able to do more and more things that improve our quality of life. More data is available than ever before, and will become even more so. This is a time when we can extract more information from data than ever before, and benefit more from it.

In different areas of business and in different institutions, new ways to collect data are continuously being created. Old documents are being digitized, new sensors count the number of cars passing along motorways and extract useful information from them, our smartphones are informing us where we are at each moment and what new opportunities are available, and our favorite social networks register to whom we are related or what things we like.

Whatever area we work in, new data is available: data on how students evaluate professors, data on the evolution of diseases and the best treatment options per patient, data on soil, humidity levels and the weather, enabling us to produce more food with better quality, data on the macro economy, our investments and stock market indicators over time, enabling fairer distribution of wealth, data on things we purchase, allowing us to purchase more effectively and at lower cost.

Students in many different domains feel the need to take advantage of the data they have. New courses on data analytics have been proposed in many different programs, from biology to information science, from engineering to economics, from social sciences to agronomy, all over the world.

The first books on data analytics that appeared some years ago were written by data scientists for other data scientists or for data science students. The majority of the people interested in these subjects were computing and statistics students. The books on data analytics were written mainly for them. Nowadays, more and more people are interested in learning data analytics. Students of economics, management, biology, medicine, sociology, engineering, and some other subjects are willing to learn about data analytics. This book

intends not only to provide a new, more friendly textbook for computing and statistics students, but also to open data analytics to those students who may know nothing about computing or statistics, but want to learn these subjects in a simple way. Those who have already studied subjects such as statistics will recognize some of the content described in this book, such as descriptive statistics. Students from computing will be familiar with a pseudocode.

After reading this book, it is not expected that you will feel like a data scientist with ability to create new methods, but it is expected that you might feel like a data analytics practitioner, able to drive a data analytics project, using the right methods to solve real problems.

*João Mendes Moreira*
*University of Porto, Porto, Portugal*

*André C. P. L. F. de Carvalho*
*University of São Paulo, São Carlos, Brazil*

*Tomáš Horváth*
*Eötvös Loránd University in Budapest*
*Pavol Jozef Šafárik University in Košice*
*October, 2017*

## Acknowledgments

The authors would like to thank Bruno Almeida Pimentel, Edésio Alcobaça Neto, Everlândio Fernandes, Victor Alexandre Padilha and Victor Hugo Barella for their useful comments.

Over the last several months, we have been in contact with several people from Wiley: Jon Gurstelle, Executive Editor on Statistics; Kathleen Pagliaro, Assistant Editor; Samantha Katherine Clarke and Kshitija Iyer, Project Editors; and Katrina Maceda, Production Editor. To all these wonderful people, we owe a deep sense of gratitude, especially now this project has been completed.

Lastly, we would like to thank our families for their constant love, support, patience, and encouragement.

J. A. T.

## Presentational Conventions

**Definition** The definitions are presented in the format shown here.

*Special sections and formats* Whenever a method is described, three different sections are presented:

- Assessing and evaluating results: how can we assess the results of a method? How to interpret them? This section is all about answering these questions.
- Setting the hyper-parameters: each method has its own hyper-parameters that must be set. This section explains how to set them.
- Advantages and disadvantages: a table summarizes the positive and negative characteristics of a given method.

## About the Companion Website

This book is accompanied by a companion website:

**www.wiley.com/go/moreira/dataanalytics**

The website includes:
- Presentation slides for instructors

# Part I

# Introductory Background

# 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.