

Data Analytics[📌] and Big Data

Soraya Sedkaoui



ISTE

WILEY

Data Analytics and Big Data

*To “Ben M’hidì”
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.

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms and licenses issued by the CLA. Enquiries concerning reproduction outside these terms should be sent to the publishers at the undermentioned address:

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

© ISTE Ltd 2018

The rights of Soraya Sedkaoui to be identified as the author of this work have been asserted by her in accordance with the Copyright, Designs and Patents Act 1988.

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	xi
Preface	xiii
Introduction	xvii
Glossary	xxi
Part 1. Towards an Understanding of Big Data: Are You Ready?	1
Chapter 1. From Data to Big Data: You Must Walk Before You Can Run	3
1.1. Introduction	3
1.2. No analytics without data	4
1.2.1. Databases	5
1.2.2. Raw data.	5
1.2.3. Text	6
1.2.4. Images, audios and videos	6
1.2.5. The Internet of Things	6
1.3. From bytes to yottabytes: the data revolution	7
1.4. Big data: definition	10
1.5. The 3Vs model	12
1.6. Why now and what does it bring?	15
1.7. Conclusions	19

Chapter 2. Big Data: A Revolution that Changes the Game	21
2.1. Introduction	21
2.2. Beyond the 3Vs	22
2.3. From understanding data to knowledge	24
2.4. Improving decision-making	27
2.5. Things to take into account	31
2.5.1. Data complexity	31
2.5.2. Data quality: look out! Not all data are the right data	32
2.5.3. What else?...Data security	33
2.6. Big data and businesses	34
2.6.1. Opportunities	34
2.6.2. Challenges	36
2.7. Conclusions	40
 Part 2. Big Data Analytics: A Compilation of Advanced Analytics Techniques that Covers a Wide Range of Data	 41
 Chapter 3. Building an Understanding of Big Data Analytics	 43
3.1. Introduction	43
3.2. Before breaking down the process... What is data analytics?	 44
3.3. Before and after big data analytics	47
3.4. Traditional versus advanced analytics: What is the difference?	 49
3.5. Advanced analytics: new paradigm	52
3.6. New statistical and computational paradigm within the big data context	 54
3.7. Conclusions	58
 Chapter 4. Why Data Analytics and When Can We Use It?	 59
4.1. Introduction	59
4.2. Understanding the changes in context	60
4.3. When real time makes the difference	63
4.4. What should data analytics address?	64
4.5. Analytics culture within companies	68
4.6. Big data analytics application: examples	71
4.7. Conclusions	75

Chapter 5. Data Analytics Process: There's Great Work Behind the Scenes	77
5.1. Introduction	77
5.2. More data, more questions for better answers	78
5.2.1. We can never say it enough: "there is no good wind for those who don't know where they are going"	78
5.2.2. Understanding the basics: identify what we already know and what we have yet to find out	79
5.2.3. Defining the tasks to be accomplished	80
5.2.4. Which technology to adopt?	80
5.2.5. Understanding data analytics is good but knowing how to use it is better! (What skills do you need?)	81
5.2.6. What does the data project cost and how will it pay off in time?	82
5.2.7. What will it mean to you once you find out?	82
5.3. Next steps: do you have an idea about a "secret sauce"?	83
5.3.1. First phase: find the data (data collection)	84
5.3.2. Second phase: construct the data (data preparation)	85
5.3.3. Third phase: go to exploration and modeling (data analysis)	85
5.3.4. Fourth phase: evaluate and interpret the results (evaluation and interpretation).	86
5.3.5. Fifth phase: transform data into actionable knowledge (deploy the model)	87
5.4. Disciplines that support the big data analytics process	88
5.4.1. Statistics	88
5.4.2. Machine learning.	88
5.4.3. Data mining.	89
5.4.4. Text mining.	90
5.4.5. Database management systems	90
5.4.6. Data streams management systems	91
5.5. Wait, it's not so simple: what to avoid when building a model?	91
5.5.1. Minimize the model error.	94
5.5.2. Maximize the likelihood of the model	95
5.5.3. What about surveys?	95
5.6. Conclusions	99

**Part 3. Data Analytics and Machine Learning:
the Relevance of Algorithms 101****Chapter 6. Machine Learning:
a Method of Data Analysis that Automates
Analytical Model Building 103**

6.1. Introduction	103
6.2. From simple descriptive analysis to predictive and prescriptive analyses: what are the different steps?	104
6.3. Artificial intelligence: algorithms and techniques	106
6.4. ML: what is it?	109
6.5. Why is it important?	113
6.6. How does ML work?	116
6.6.1. Definition of the business need (problem statement) and its formalization	117
6.6.2. Collection and preparation of the useful data that will be used to meet this need.	117
6.6.3. Test the performance of the obtained model.	118
6.6.4. Optimization and production start.	118
6.7. Data scientist: the new alchemist.	120
6.8. Conclusion	122

**Chapter 7. Supervised versus Unsupervised Algorithms:
a Guided Tour 123**

7.1. Introduction	123
7.2. Supervised and unsupervised learning	124
7.2.1. Supervised learning: predict, predict and predict!	124
7.2.2. Unsupervised learning: go to profiles search!	127
7.3. Regression versus classification	129
7.3.1. Regression.	130
7.3.2. Classification	133
7.4. Clustering gathers data.	141
7.4.1. What good could it serve?	141
7.4.2. Principle of clustering algorithms	144
7.4.3. Partitioning your data by using the K-means algorithm	148
7.5. Conclusion	151

Chapter 8. Applications and Examples 153

8.1. Introduction	153
8.2. Which algorithm to use?	153

8.2.1. Supervised or unsupervised algorithm: in which case do we use each one?	154
8.2.2. What about other ML algorithms?	157
8.3. The duo big data/ML: examples of use	161
8.3.1. Netflix: show me what you are looking at and I'll personalize what you like	162
8.3.2. Amazon: when AI comes into your everyday life	165
8.3.3. And more: proof that data are a source of creativity	168
8.4. Conclusions	171
Bibliography	173
Index	181

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

⊙○○◌◌◌◌◌ ◊◊∧ℝ◌◌◌◌◌

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.