Premier Reference Source

Big Data Analytics for Entrepreneurial Success

Soraya Sedkaoui

Big Data Analytics for Entrepreneurial Success

Soraya Sedkaoui
Khemis Miliana University, Algeria & SRY Consulting Montpellier, France

A volume in the Advances in Business Information Systems and Analytics (ABISA) Book Series

Published in the United States of America by
IGI Global
Business Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA, USA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Names: Sedkaoui, Soraya, author.
Title: Big data analytics for entrepreneurial success / by Soraya Sedkaoui.
Description: Hershey, PA : Business Science Reference, [2019]
Identifiers: LCCN 2018031991| ISBN 9781522576099 (hardcover) | ISBN 9781522576105 (ebook)
Subjects: LCSH: Management--Statistical methods. | Business planning--Statistical methods. | Big data. | Entrepreneurship.
Classification: LCC HD30.215 .S43 2019 | DDC 658/.0557--dc23 LC record available at https://lccn.loc.gov/2018031991

This book is published in the IGI Global book series Advances in Business Information Systems and Analytics (ABISA) (ISSN: 2327-3275; eISSN: 2327-3283)

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Advances in Business Information Systems and Analytics (ABISA) Book Series

ISSN:2327-3275
EISSN:2327-3283

Editor-in-Chief: Madjid Tavana, La Salle University, USA

MISSION

The successful development and management of information systems and business analytics is crucial to the success of an organization. New technological developments and methods for data analysis have allowed organizations to not only improve their processes and allow for greater productivity, but have also provided businesses with a venue through which to cut costs, plan for the future, and maintain competitive advantage in the information age.

The **Advances in Business Information Systems and Analytics (ABISA) Book Series** aims to present diverse and timely research in the development, deployment, and management of business information systems and business analytics for continued organizational development and improved business value.

COVERAGE

- Data Management
- Geo-BIS
- Algorithms
- Business Information Security
- Business Intelligence
- Business Systems Engineering
- Decision Support Systems
- Business Process Management
- Data Governance
- Big Data

IGI Global is currently accepting manuscripts for publication within this series. To submit a proposal for a volume in this series, please contact our Acquisition Editors at Acquisitions@igi-global.com or visit: http://www.igi-global.com/publish/.

The Advances in Business Information Systems and Analytics (ABISA) Book Series (ISSN 2327-3275) is published by IGI Global, 701 E. Chocolate Avenue, Hershey, PA 17033-1240, USA, www.igi-global.com. This series is composed of titles available for purchase individually; each title is edited to be contextually exclusive from any other title within the series. For pricing and ordering information please visit http://www.igi-global.com/book-series/advances-business-information-systems-analytics/37155. Postmaster: Send all address changes to above address.

Titles in this Series

For a list of additional titles in this series, please visit:
https://www.igi-global.com/book-series/advances-business-information-systems-analytics/37155

Sentiment Analysis and Knowledge Discovery in Contemporary Business
Dharmendra Singh Rajput (VIT University, India) Ramjeevan Singh Thakur (Maulana Azad National Institute of Technology, India) and S. Muzamil Basha (VIT University, India)
Business Science Reference • ©2019 • 333pp • H/C (ISBN: 9781522549994) • US $215.00

Law, Ethics, and Integrity in the Sports Industry
Konstantinos Margaritis (University of Crete, Greece)
Business Science Reference • ©2019 • 307pp • H/C (ISBN: 9781522553878) • US $195.00

Institutional and Organizational Transformations in the Robotic Era Emerging Research ...
Albena Antonova (Sofia University, Bulgaria)
Business Science Reference • ©2019 • 178pp • H/C (ISBN: 9781522562702) • US $155.00

Utilizing Big Data Paradigms for Business Intelligence
Jérôme Darmont (Université Lumière Lyon 2, France) and Sabine Loudcher (Université Lumière Lyon 2, France)
Business Science Reference • ©2019 • 313pp • H/C (ISBN: 9781522549635) • US $210.00

Handbook of Research on Expanding Business Opportunities With Information Systems ...
George Leal Jamil (Informações em Rede Consultoria e Treinamento Ltda, Brazil)
Business Science Reference • ©2019 • 455pp • H/C (ISBN: 9781522562252) • US $245.00

Qualitative Techniques for Workplace Data Analysis
Manish Gupta (IFHE University, India) Musarrat Shaheen (IFHE University, India) and K. Prathap Reddy (IFHE University, India)
Business Science Reference • ©2019 • 317pp • H/C (ISBN: 9781522553663) • US $215.00

Machine Learning Techniques for Improved Business Analytics
Dileep Kumar G. (Adama Science and Technology University, Ethiopia)
Business Science Reference • ©2019 • 286pp • H/C (ISBN: 9781522535348) • US $195.00

For an entire list of titles in this series, please visit:
https://www.igi-global.com/book-series/advances-business-information-systems-analytics/37155

701 East Chocolate Avenue, Hershey, PA 17033, USA
Tel: 717-533-8845 x100 • Fax: 717-533-8661
E-Mail: cust@igi-global.com • www.igi-global.com

To the memory of my father:
In my heart, you will always be loved and remembered.

Table of Contents

Section 3
Entrepreneur! Welcome to Your Data-Driven Universe

Preface

In an unstable and rapidly changing business context, I dedicate this book to all the managers, entrepreneurs, and everyone who continues to believe every day in a better future, and who contributes to its creation.

Today, in many ways the situation has changed. The global market has become a reality, investment and support structures have emerged. And new digital technologies offer the entrepreneur exceptional opportunities for technical and commercial development and financing.

An entrepreneur nowadays can finance the first stages of his project and find his first clients via crowdfunding. He can develop new software for a fraction of what it would have cost 10 to 15 years ago. He can quickly implement digital tools to directly access a market that extends well beyond our borders.

Entrepreneurial practices are then constantly evolving. It is difficult to seize them. We have to notice that, the entrepreneurial practices of today are not the same as those of yesterday and they are different from those of tomorrow.

While most reflections in the field of entrepreneurship attempt to answer the question of: "what is entrepreneurship?" by offering contemporary or historical descriptions of this concept (Chapter 7), this book takes an original look at the entrepreneurship of tomorrow, a look of the near future: that related to the emergence of big data analytics, data science, and machine learning. These concepts have been widely discussed, not just throughout this book but virtually around the world.

Big data here, big data there, is the expression of the moment when we evoke the digital revolution. This has happened not only because of the increase in computing power and the availability of large amounts of data but also through the adoption of new analytics algorithms. There is no doubt that this trend will accelerate, fueled by scientific and technological advances that will not stop.

The grouping of a multitude of data is today at the center of all the concerns, due to the consequences that can result from it. The power of the good use of this data is just extraordinary in many areas. In 2018, big data is no longer a concept, it is a reality. But exploit and process these mass data produced in real-time and in different types, is not yet within the reach of everyone.

To get the basics right, a scientific and technical cocktail is needed: statistical and mathematical bases, on the one hand, computer and programming skills on the other hand. Without neglecting real analytical and creative skills to understand application issues and translates business problems into value.

This original book, in its style and content, can help you to develop them. After reading it, you will have a solid knowledge of the big data analysis methods and machine learning algorithms (Chapter 6) that are most common today in the majority of applications, from the simplest one to the most advanced. You will know the principles, exposed in an accessible and relaxed way.

This book will help you to familiarize yourself with the use of data and the analytics applications (Chapters 4 and 5), that are at the heart of big data and data science universe, and acquire the skills and knowledge needed to act as a data scientist.

As such, this book clarifies and details many aspects that any modern entrepreneur or data entrepreneur must master (Chapter 8). It's a must-have for any professional, student or curious people interested in the world of data as a whole. Without being dogmatic, it is pragmatic, educational and provides clarity on the data-driven solutions currently in force. Through this book, the author aims to make you aware of this new reality of digital change and help you to better prepare for it.

Faced with this universe, this book aims to help people who want to join the big data analytics arena to:

- Implement a robust data strategy and industrialize a set of "data-driven" business use cases;
- Have a holistic view of the available data and how to collect, store and analyze them to generate value;
- Deepen their knowledge about the main technologies of big data (Hadoop, MapReduce ...);
- Understand the basics of data analytics applications in business (regression, classification, clustering ...).

In a resolutely practical and data-driven project universe, many examples illustrate the theory (Chapter 9), and demystify the buzzwords that invade

the discussions around big data (Chapters 1 and 2) and bring, of course, the reader from confusion to clarity! In parallel with this, we tried as much as possible to open the horizons and to sharpen the curiosity of the reader with various examples (Chapter 3).

One of the other strengths of this book is that it relies on my concrete experience. Through my field approach to big data analytics, I can share real examples and practical implementation tips. I believe that every data byte has something to tell and that behind billions of data bytes there are enormous underlying messages to be revealed.

I have been passionate for over 10 years now by this field in order to understand these messages and design algorithms and methods to analyze them. These years have been dotted with fabulous ideas.

I am convinced that these details make the difference between the person who has only taken theoretical courses in big data analytics and data science and the true practitioners who put their hands in the process and manipulate the data rigorously. In this sense, this book is a good complement to more theoretical academic works on the subject. It does not claim to be a rigid and closed reference book in this universe; on the contrary, it is an open window to the vast world of data analysis. So, at the end of this book, if you feel that some methods are missing, that is because you are forging your own personality as a "data scientist".

WHY THIS BOOK?

The big data analytics, objective of this book, is a still young discipline at the crossroads of several domains. It now represents an extension of data analysis to the business applications and other technical fields that are needed to harvest, manipulate and leverage the data available today.

A good data entrepreneur must be able to navigate between these different disciplines that include: statistics, algorithms, IT, without a priori theoretical. What is important is the ability to find an adequate answer to a given problem. In this sense, his main quality will be his ability to understand his field of action and find the best solution among the many technical choices: technology, platforms, software…, and theoretical: methods and algorithms, possible, given the time and budget constraints.

In the context of this book, I cannot, of course, address the subject exhaustively. I have delimited its perimeter as follows. First of all, I considered that the primary role of the data entrepreneur is to know how to extract the

knowledge from the available data, to answer a given problem. I will, therefore, focus on the issues of data analysis, and not on the technical constraints of its implementation. Then, I decided to approach this question mainly from analytics angle. Because the usual analytical methods and their applications are already the objects of many publications and I would not have much more to bring moreover. Also, the analytics approach is currently booming, especially in the field of big data and machine learning.

I strongly believe in these practices and wish to contribute to their dissemination. I believe this is where the future of the entrepreneur lies, and it is not for nothing that Harvard Business ranked a data scientist as: "the sexy new job of the 21st century".

Finally, there is still little work addressing the issue of big data analytics from a practical angle for entrepreneurs, with a pedagogical explanation of the operation of its main methods and concrete demonstrations of their use. It is this emptiness that this book hopes to fill, in order to help practitioners: entrepreneurs, data scientists, students and all other interested people, to better leverage the power of analytics.

As a "Data analytics Bible" for an entrepreneur, this book proposes in a new way a robust method to exploit the full potential of the data. It will enable modern entrepreneurs to:

- Understand the mechanisms for data collection, storage, and reconciliation;
- Master the main tools of data analytics;
- To demystify the analytical and statistical concepts applied to the business (clustering, regression, predictive method ...).

The point is not to present the current trends in entrepreneurship, but rather to identify emerging perspectives. This temporal dimension of entrepreneurship is as important as contemporary and historical descriptions. This can be understood by the desire to capture and crystallize the evolution of entrepreneurial practices in order to attract the attention of practitioners, students, but also entrepreneurs about the importance of the data available today, because there is a plenty of opportunities to be seized, these are all avenues of action to consider, to discuss, to try and to undertake.

The reader will, by reading each chapter of this book, have all the keys to concretely tackle a big data project: analyze a problem from the "analytical" point of view, apply a model cleverly and understand how and why to create

it. And because data analytics is learned primarily through practice, the book will integrate concrete examples from cases encountered during the author's professional career on the big data battlefields.

WHO ARE YOU?

Who are you to focus on the business implications and value of big data analytics? Readers of my previous works on analytics were generally ambitious professionals who thought that data analytics and technology could change business rules and practices. I suspect that you have the same attributes.

The most interested are often those whose functions involve a lot of data, such as marketing, the supply chain, finance and human resources. People, who set up big data, like the IT department, also want to deepen their knowledge of the concept. Interest is even greater in industries that are already heavily data-driven, such as online businesses, or those that have the potential to transform themselves - potentially any industry, but especially those with a large number of customer data - such as retail, tourism and transportation, telecommunications, media and entertainment, and more.

If you are still a student and planning a career, or a job, in big data, congratulate yourself for your insight. This sector is likely to remain in full swing for many years.

Globally, this book is for entrepreneurs, but also for anyone curious to have an overview about the state of the art of big data analytics and its various applications in business context. It is 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.

In concrete terms, this book aims to help the reader to answer the following question: "If I have to tackle a big data project, how should I do it?". Perhaps that is why you are reading this book: you want to help your business answer such questions and probably make your career better. You have chosen the right starting point!

This book was put to answer this question. The author will tackle this question by gathering a collection of chapters with a high scientific quality; rich in knowledge, methods, and concepts, in order to, carefully, put it as simple as possible to help you drawing your proper path in this universe.

For this, a minimum of the theory is, of course, necessary to choose the right algorithm, to understand how it works and what to expect. But, also do not forget that when working with data the results can be used to make decisions and generate business value, so a bit of theoretical rigor is also needed.

A minimum of statistical and computing knowledge are also required to read this book, even if I have tried to make it as accessible and educational as possible. In order to help those people who are interested in developing a board picture of the current context characterized by the 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.

Beyond the state of the art, this book will also interest anyone who wants to implement a big data project. Knowing the algorithms is good but what you can do with it is better. For this, a set of tips must be known to enrich, manipulate and successfully exploit the data and interpret the results. These are the tricks of practitioners, resulting from my data scientist experiences at SRY Consulting (Montpellier).

HOW TO READ THIS BOOK

Data is a central topic because it gives business professionals the opportunity to increase the potential of their strategies. Enabling data and turning it into Insights is a high-performance business strategy, but one that may seem complex or even scary. However, it is not. You can easily evolve your analytical capabilities and create opportunities at every stage of your analytics process.

Dive into the different chapters of this book and discover how you can successfully, step by step, launch your big data project and diversify your menu of big data possibilities, understand the big data universe and begin a more efficient analytics process. The profusion of today's data contains answers to questions that no one has thought of asking yet. This book will guide you to get these answers, through nine chapters.

If you are a novice in this field, I recommend you to read this book linearly. In this case, concepts related to big data and analytics will be gradually introduced, with increasing complexity. As much as possible we have directed each chapter so that it can be read independently of the others. In order to

organize your reading, know that the chapters of this book are articulated according to the three main sections. A brief summary of each section and chapters are provided below.

Section 1: Big Data and the Business Context – Mania vs. Phobia

The first section will interest any type of reader because it discusses the general context of the big data universe and presents the corresponding state-of-the-art. It offers, through three chapters, the main principles of the big data context. It will allow you to know in what state of mind you are situated and what are the main reflexes to acquire to tackle a big data problem. This first section introduces the reader to the several Vs related to big data, before explaining its challenges and opportunities and how the digital age has led to changes the way data is collected, stored, analyzed, visualized and protected. It, also, highlights the importance of data in business and how it can increase its efficiency. This section gives a platform to proceed to different big data related concepts and how this phenomenon is changing business opportunities, by summarizing a great set of business examples. It examines the various ways that big data can benefit businesses.

Chapter 1, "Big Data, Who Are You?" gives an overview of the basic concepts of big data and discusses the several main opportunities and challenges related to the "3Vs" phenomenon in the business context.

Chapter 2, "What the 3Vs Acronym Didn't Put Into Perspective," advocates for the other side of the big data phenomenon that lies behind the "3 Vs" and can interest the reader (smart data, data quality, privacy …).

Chapter 3, "Big Data Applications in Business," presents a panorama of big data applications. These applications come from the business playground and show the interest that big companies' accord to the big data universe, and how it allowed them to achieve their business strategies and boost their value creation chain.

Section 2: The Hello World of Big Data Analytics

The second section is addressed 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. This section will provide a complete overview related to the data analytics process.

So, don't worry, because even if you are completely new to big data universe, analytics techniques and the machine learning algorithms application, this section will illustrate the power of advanced analytics and its wide range applications by showing how it can be applied to solve fundamental data analysis tasks. This section will explain some of the new methods and algorithms required to exploit the large volume of available, which can change the way you think about the analytical process.

Chapter 4, "First of All, Understand Data Analytics Context and Changes," covers the most basic form of data analysis, from descriptive to predictive to prospective analysis. This chapter is also an opportunity for the reader to discover the challenges that data analysis should address in the big data age.

In Chapter 5, "Understanding Data Analytics Is Good but Knowing How to Use It Is Better!" the author goes over the different steps involved in the process of data analysis, as well as take the reader through the various tools, advanced technology and skills required.

Chapter 6, "Techniques and Methods That Help to Make Big Data the Simplest Recipe for Success," demonstrates that the field of big data analytics is so vast. Coupled with the algorithms of machine learning, many methods can be used to better analyze data and reap the maximum benefits.

Section 3: Entrepreneur! Welcome to Your Data-Driven Universe

The third and last section will be dedicated to the entrepreneurs who want to embrace this universe and launch their big data project. It gives a set of keys to anyone wishing to generate value from data using analytics tools and methods. In the form of a bestiary, this section details how an entrepreneur can convert data into value. It goes beyond the simple enumeration of algorithms. As we use them, we introduce the fundamental principles of data analytics. In addition to purely operational aspects, this section will discuss a number of fundamental concepts of data analytics, such as the definition of the problem to be addressed, the treatment of missing values ... And because the knowledge of algorithms is not enough to become a "data entrepreneur", this section will show concretely how to implement a big data project from A to Z. To better understand the principles evoked throughout the book, practical examples have been introduced and detailed, in order to help the curious and willing reader to go further in this area and deepen these aspects with the provided examples.

The Chapter 7, "Entrepreneurship and Big Data," is quite detailed and potentially useful to the reader to receive a total systems approach around the topic. In order to improve entrepreneurial competitiveness and performance, the entrepreneur must embrace advancements in digitalization. Successful implementation of data-driven culture is a huge factor to successfully conduct the big data project.

The Chapter 8, "Plan and Rules for Data Analysis Success: A Roadmap," contains a workflow by drawing a successful way through which an entrepreneur can improve the effectiveness of his big data project. It can be considered as a reference because it provides a guide to prepare a roadmap and draw the key elements of this project and how to address its challenges.

In Chapter 9, "Big Data Analytics in Action: Examples," the reader can find different examples where several data analysis methods have been used. These examples can help to understand how to manage the available data and how to apply algorithms using software and big data technology like MapReduce.

In conclusion, my ambition to write this book was to make it one of the first foundation references to the big data analytics in the entrepreneurial context. I hope that will enjoy this book and find it valuable.

Good reading!

Soraya Sedkaoui
University of Khemis Miliana, Algeria & SRY Consulting Montpellier, France

Acknowledgment

He who does not thank people does not thank God. (Prophet Mohamed, PBUH)

I could not have written this book if I had not learned data analytics, and I probably would never have done work in this exciting area without the influence of Professor Hans-Werner Gottinger.

He introduced me to a host of new concepts, put me in contact with extraordinary people and made me aware of my imperfections to the point that I was able to realize several works and write books to compensate. Special thanks, as the editor of this book, to Hans (again him) who pointed out some tips on how to work with data. He helped me to better structure my ideas through his enlightened remarks. Thank you for giving me your support to carry out this project.

Above all, I owe a huge thank to my family. The only thing harder than writing a book is to live with someone who writes a book. I could never have finished this book without their support.

Many other people have also indirectly but strongly contributed to the content of this book: my colleagues and my friends with whom I constantly exchanged new ideas.

Of course, I thank IGI Global for their support and trust in writing this book, especially Marianne Caesar, Courtney Tychinski, and Josephine Dadeboe for their support during all along the editing process.

Introduction

They say one of the best ways to learn is by doing. So, surely another great way is by reading.

Like every other production factor, our modern business activity could not happen without 'data'. This new asset is on the lips of all managers and decision-maker, and made the headlines and leads entire conferences. Collect, analyze, and interpret data, become an intrinsic and essential part of business operations to successfully conduct many activities. This requires a new way of thinking in how data is stored, processed, analyzed, and protected.

This book, *Big Data Analytics for Entrepreneurial Success*, as the title states, provides a good source of emerging techniques, methods, tools, way, process and applications of this fascinating field in the implementation of new business ideas. It is a pivotal reference source that provides vital bases on how to extract value from the big amount of data and address its challenges. It is a vital publication that examines this paradigm that has influenced major changes in many business aspects through analytics approach.

You are probably thinking of "here is yet another story that thinks to be original about big data analytics and how it is supposed to change my business". But, let me, through this book, introduce you to some of my experiences before deciding if you do not actually want to hear anything related to the big data.

My belief is that large companies should not be the only ones to benefit from the technological and analytical advances and the new perspectives they bring. Entrepreneurs or small businesses are also concerned. In my opinion, we must remove the "big" superfluous and keep only the "data" relevant to this phenomenon. This is important to address the different ways of analytics using with more insight, to improve your business activity by bringing more pragmatism around the famous "big data".

Big data analytics does not necessarily mean complex solutions and heavy and time-consuming implementation projects. This is why we should not hesitate to put things in perspective and adopt an intelligent approach compared to larger budget projects more commonly.

You have to know that the solution is always hidden in the data that needs only to be extracted and understood properly. The first step to being successful is the proper understanding of the problem. Do you want to launch a big data project? So before making plans, start with understanding. As a result, an effective solution can be sought. It is obvious that the more you work on analytics-related projects, the better you become to generate good solutions.

You may say: "I'm a dreamer and I can do it". Yes! You can join this analytics arena, but you have to know that you are not the only one who thinks so. In addition, from the dream to the reality, the road needs many aspects. So, it is important to keep in mind that it is not just about fees or the nature of technology adopted. When implementing such a project, the most important thing is how to generate more value.

Is big data important to you and your business? So, you have to act. You need to identify the aspects that can be applied to your business context and work on them. You must adopt approaches or create solutions that generate value to make big data operational. Change the IT architecture is also needed. If you get into all these actions, you can then do it better and successfully.

Beyond these aspects, appears a wider debate, "Big data: culture or technology?" Indeed, setting up a big data project requires setting up new tools and new processes. It requires also rethinking the organization of the business activities, even considering outsourcing, restructuring the management and reviewing the required skills, etc. In short, big data implies a real transformation of the business and the setting up of a new culture, essentially the data-driven culture.

So, these large amounts of data must have meaning for you, and in any case, you must first be convinced! If you decide to go further into the big data universe, you have to discover the most economical and affordable ways.

You have time constraints and priorities, but you need to be ready to mobilize resources around your big data project, so start with an exploratory project to see the new possibilities. Businesses should embark on this path; at least discussed at the highest level the place that can take big data.

Rather than announce: "I am embarking on a big data project", it is probably more constructive to say: "I will analyze the data to better understand and boost my project and generate business value". In addition to the clarity of

your intentions and your strategy, this approach avoids endless discussions about the size, big or small, of the data involved.

In what follows I expose my insights on the reasons that push you to plow in this universe. The rest of this book will explain why big data and analytics are newsworthy. More data and more technologies, what else is needed to make sure that your big data project is not just a buzzword?

As we will see throughout this book, it is the people who will give big data its real measure. Your way of doing things is the main factor to successfully guide your big data project. So you need to understand before undertake, and always undertake to innovate.

Soraya Sedkaoui
University of Khemis Miliana, Algeria & SRY Consulting Montpellier, France

Section 1
Big Data and the Business Context:
Mania vs. Phobia

Big data is watching you.
But don't be afraid! …Relax, take control, and read this section.

Having an overview of your data is not an easy task because there is no classification that refers to the different issues and applications of big data in the business context. To avoid getting lost in the labyrinth of big data, it is often beneficial to focus first on the useful data: by understanding its nature, its opportunities, and its challenges (Chapter 1); to ensure the quality and accessibility of data and build (Chapter 2); from the foundation of the cases of uses (Chapter 3). This section is quite theoretical but necessary to understand the big data universes. It gathers a rich collection of many related concepts and discusses the open issues and challenges related to the big data revolution in the business context. Through three chapters, this section introduces this book by giving an overview of basic notions, and presents current and the most probable applications of big data, as well as actual business examples.

Chapter 1
Big Data, Who Are You?

ABSTRACT

When you hear "big data" probably you think like most, to a rather heavy file that includes diverse information, or a real iceberg that hides a large portion of its real mass. But we have not to think so about big data. Yes! You can worry about it. But you can very well choose to see it as a potential which allows you to create value. I would say even more that you must choose to take advantage of this large amount of data! Data or big data are primarily a way to segment your target. It is a collection of information that is analyzed, processed, and arranged to be profitable. So, do not let data phobia hold you back, just keep calm because this first chapter gives you an overview of what the concept big data encompasses, and you will realize by yourself that it is an exciting field.

INTRODUCTION

Data! data! data!, he cried impatiently. 'I can't make bricks without clay'

Sir Arthur Conan Doyle (1892). (Adventures of Sherlock Holmes, p.289)

You have probably seen it yourself in magazines, on TV or even heard by your friends, the term "Big Data" is now more than ever a fashion word. But what lies behind this little vague concept? An explanation is needed. The concepts behind big data are actually nothing new because data existing over the time, but what makes it so important is the rapid rate and different types

DOI: 10.4018/978-1-5225-7609-9.ch001

in which it is produced in recent time. This brings us to say that data become big: it is "big data!". In another word, big data is the new form of data which seemed to come back to presenting old wine in a new bottle.

Data has grown from kilobytes (KB) to petabytes (PB). This huge amount of data is referred to as big data and requires advanced tools and software for processing, analyzing and storing purposes. This phenomenon has radically changed the way data are collected and analyzed since it introduces new issues concerning volume, velocity and the variety of data.

I already hear many people questioning the impact of big data and the risk of restricting the creative process. Think again! We should rather see big data as a support for creation. The data will never dictate how to produce creative content, but it can clearly indicate what it should contain in order to capture the interest of events and generate value. People in the business context no longer have to depend on their intuition to ensure the success of their creations. They now have at their disposal a tool that supports them, and which has a much weaker bias.

So should we be afraid of big data or should we exploit it? Between those who think that it is a time saver and a factor of considerable growth and those who are against the exploitation of the data because they consider that it is a violation of their private life; the answer can be found and discovered by yourself by reading this chapter.

Because the new business revolution is based on these data, these billions of tiny pieces of information that allows reinventing, to adapt and to redraw all businesses and services, provided that we can draw useful information and create a minimum of readability in this galaxy of data. This is the real challenge.

Big data is the revolutionary word in today's world because of its influence on several domains. It is said everywhere that it is the future and that you have to start exploring big data as soon as possible. All right, but where do you start? What to learn? How to extract value from a large volume of available data? Which technology must you adopt? And so on.

Rather than propose an endless inventory types formations, with titles that will maintain this sense of key buzzwords, we preferred to discuss the big data overview: How and why did we get there?

It is by having a better vision of what big data is that you will deduce by yourself what you must look for to become an ace of the subject. Especially, what you need to learn by yourself. This chapter will clarify what big data really is, where it comes from and how it is processed.

NOT ONLY BIG BUT ALL ABOUT 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 of the business activities are associated with the use, the understanding and the exploiting of data (Sedkaoui, 2018a). "Data is everywhere", "data is the new oil", "digital oil", "data is power" … words are diverse and do not miss to describe the importance of 'data'.

Real atom, data are, over the time, at the heart of thinking about new business strategies, and often the fantasy of all business models for its value. So why suddenly the term 'data' is in every conversation and many research are published on the subject throughout the world? What has changed so much and justifies such enthusiasm sometimes verging on collective madness?

Digitalization of Everyday Life

Since you have read this chapter, thousands of tweets have been exchanged, millions of queries have been analyzed by Google, millions of likes on Facebook have been attributed, more than one hundred hours of new YouTube videos have been uploaded, and several Netflix videos launched! In total in less than a minute reading a huge amount of data was created.

Big data is often related to IT solutions for data processing and storage; hence the feeling of "too technical for me!" is it not? But what we don't integrate is that our daily lives are centered and guided by these data. Yes! Today each one is, in one way or another, the producer and consumer of data. But how is this possible? Just go behind the scenes of all of these connected objects that we have become dependent on smartphones, computers, tablets, smart watches, GPS, smart car ... and the various available applications.

This massive influx of data has completely changed the paradigm that prevailed in the 1980s. The differentiating factor of a society today does not lie in having data than being able to analyze them and turn them into information! Data is abundant, so what is useful and rare is the ability to exploit it and make it operable.

Increase in Calculation Power

Being able to process large volumes of data quickly and cheaply is the second major change in data growth. This is mainly related to:

- **Moore's Law:** Moore's law states that the number of transistors in a dense integrated circuit doubles approximately every eighteen months, without increasing cost.
- **Open Source:** This refers to free software built by "non-profit communities". Most big data architectures are composed of such open source software, the most famous of which is Hadoop.
- **Cloud:** The cloud has drastically reduced the cost of data processing. In the collaborative economy model, for example, it becomes possible for any start-up to rent in real-time as many machines as necessary.

The Growth of Storage Capacity

Big data is a board term for data sets so large and complex that traditional data processing is inadequate. At first, the rapid multiplication of data collected, stored and correlated by companies does not seem to pose a major problem. The capacity of storage media is increasing, while the price per gigabyte is decreasing. Processor performance is growing exponentially as if hard drives were applying their own version of Moore's Law.

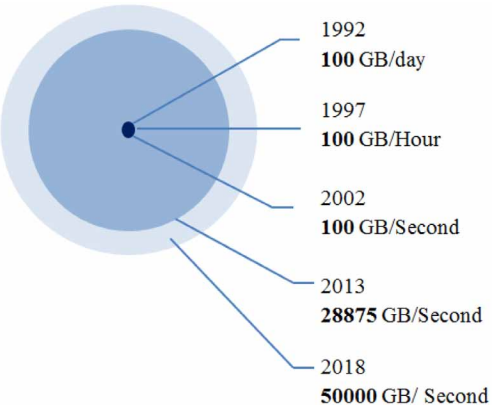
However, not only does the storage capacity curve increase less than that of the processors (see figure 2 in chapter 2, which summarize costs of storage and data availability), but its progression is also insufficient to cope with the explosion of data. In this context, Big Data pushes the limits of Moore's Law, since over 15 years, the storage capacity has increased from 40 MB to 750 GB (x 60000) when, at the same time, the bit rate has simply been multiplied by 100.

Business Changes

More scientific, more technological, modern business is also more creative. The appearance of advanced analysis tools releases the operational potential of the business. Big data has become one of the most exciting fields of our time, which opens the way to new opportunities that have significantly changed the business playground.

Starting from these circumstances, the concept of big data is born. It covers three dimensions: the exploitation of the exponential flow of data generated (Volume), coming from different sources and in different formats (Variety) and crossed most often in real-time (Velocity). The financial returns from their large-scale analysis could be in the billions of dollars.

Figure 1. The growth of storage capacity



The awareness of the critical importance of the data for the business has taken place. It is now a matter of making the most of this data, for operational projects (revenue growth, cost reduction, etc.) but also and above of all for a strategic project, such as successful in its digital transformation.

Data is the lever that achieves this goal: thanks to the new storage and algorithmic processing, data analysis offers many new uses. And, the most common uses concern the refinement of commercial proposals, the optimization of its activity, and the creation of new services... These are all areas that are transforming the way the business acts.

Data existing over the time, it's not new, but what makes it so important is the rapid rate and different types in which it's produced in recent times, or what brings us to turn: *"From data to Big data"*. This brings to the metaphor of *Dan Ariely*, professor of psychology and economics, which expresses the vagueness that surrounds today the "big data" age.

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it.

Working with big data requires a set of IT tools with a new form of integration to reveal insight from the available data that are diverse, complex, and produced in real-time. Moreover, many authors point out that big data cannot be analyzed with traditional methods; a development of efficient methods and technologies is therefore needed (Chen et al., 2014). In this

context, understand the development of the related concept, methods, and technology is necessary. 'Big data chronology', that's what will be detailed in the following point.

BIG DATA CHRONOLOGY

The digital revolution we are going through is a great opportunity, and now everything is playing out. Never companies had received such a lever for growth and competitiveness, whatever their sector, activity, size, location... This revolution has become one of the major strategic issues for companies, whether big companies or small and medium enterprises (SMEs). It focuses more on upgrading companies for the systematic use of digital technologies, web methods and uses, both internally and externally. The volume of data produced has increased significantly in effect, as its processing capacity. This widespread production of data has resulted in the 'data revolution' or the age of 'big data' (Sedkaoui, 2018b).

To better understand big data, its challenges, its opportunities, its promises, and applications, as well as advanced analytics models and mechanisms, it's necessary to know who coined big data and why.

We have the opportunity to come back, in this first chapter, at the chronological presentation of key figures in the development of big data: From *John Mashey* who, in 1989, was the first man to use the assembly of words that defines our subject, to the creation of the first version of *Hadoop* by *Doug Cutting* in the early 2000s that symbolizes the true birth of Big Data technologies, through *Sergei Brin* and *Larry Page*, founders of the overpowered Google, a member firm of the restricted big four of the technology market.

It is in the archives of the digital library of the '*Association for Computing Machinery'* that we find the term big data, used for the first time in 1997. At the time, its meaning is already close to that of today since it designated 'large data sets'.

With the ever-increasing volume of data, it was necessary to rethink the storage and the processing of this volume, to keep extracting information quickly. Thus *Doug Laney* enunciated a few years later the 3V rules to standardize a definition of big data. A fourth 'V' has even been added to a posteriori, that refers to 'Value' it is related to the companies' objective or the benefit generated from data, especially by making sense out of it (Sedkaoui and Monino, 2016).

Table 1. Data history

Age	Usage	Methods
Antiquity	The data was used to identify assets and people, to raise armies and taxes. The traders start counting their goods and currencies to manage the activities.	Census Accounting
Industrial age	With the industrial revolution, the notion of statistical law appears, and with it the paradigm of polls and that of modernization, which make it possible to observe phenomena, but also to explain and predict them. Companies move from simple accounting to management control, performance management, and business intelligence.	Management control Risk measurement Performance management Reporting Business intelligence
Digital age	Data becomes a key element of transformation and innovation. With the digital revolution, statistics become data science. And the data is big data which becomes an important factor in the transformation of the business.	IoT Analytics Data analytics Machine learning Algorithms Deep learning

If we look at the chronology of the events illustrated in Table 2 that generated big data as we know it today, we can see that this genesis goes back quite a long way and that it took the ideas to convince and that solutions ripen to get there. So, it is to say that the challenges of this Mount Everest of data are far from new but can be explained by the 3 Vs (volume, variety, and velocity) that characterize any big data definition. It is often to address big data paradigms by addressing many different "Vs" of big data.

THE 3 V'S OF BIG DATA

Big data is now a major issue for modern economies. For companies that implement strategies around it, these strategies are sources of revenue, loyalty, and innovation. The concept of big data has been theorized by Gartner (2013), and defined as follows:

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

Research firm McKinsey (2011) 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

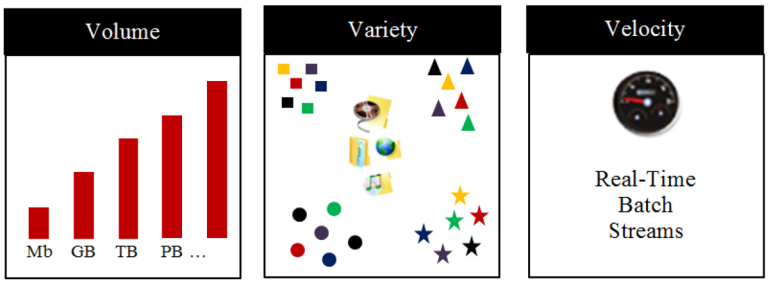
a dataset needs to be in order to be considered big data — i.e., we don't define big data in terms of being larger than a certain number of terabytes (thousands of gigabytes). We assume that, as technology advances over time, the size of datasets that qualify as big data will also increase. Also note 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).

The relativity of the word 'big' can make the definition of 'big data', in some way, complex but a generally accepted fact is that for data to qualify as big data, as can be noticed from the two previous definitions, it must have either or all of three characteristics namely: *volume* (how much?), *velocity*

Table 2. Events that generated big data

Year	Event
1997	First appearance of the term big data (NASA) to describe the challenge of working with large volumes of unstructured data.
1998	First NoSQL database (and use of the term) by Carlo Strozzi
2000	Neo4j (NoSQL base oriented graphs)
2001	First Definition of big data (Volume / Variety / Velocity) by Gartner
2005	• Birth of Hadoop • Birth of CouchDB (document-based NoSQL database)
2006	Google BigTable Publication
2007	• HBase birth (document-based basis) • Amazon Dynamo-based column-based publishing • MongoDB database (document-based database) • Creation company 10Gen (become MongoDB)
2008	• Birth of Cassandra (partitioned NoSQL database) • Creation company Cloudera • Hadoop breaks the record "Terabyte spell Benchmark"
2009	• Birth of Flink • Redis Key / Value Base • Creation of the company MapR • Birth of Mesos (resource management) • Birth of Spark (data analysis)
2010	Creation of DataStax (Cassandra)
2011	Creation of Hortonworks
2012	YARN to replace Hadoop v1
2013	Creation of DataBricks (Spark)
2014	Spark breaks the record "Terabyte spell Benchmark"

Figure 2. The 3 V's of big data



(at what speed?) and *variety* (how diverse?) (Zikopoulos and Eaton, 2011; Sedkaoui, 2018a). Typically, the '3Vs' are used to characterize the key properties of big data (see Figure 2).

Big data implies an enormous *volume* of data, which is the most associated 'V' with this phenomenon. What we are talking about here is the amounts of data reaching almost unimaginable proportions. However, this first V is the least operating and most variable depending on the sector and the organization. Today we talk about storing and analyzing Exabyte (10^{18}) or even Zettabyte (10^{21}), whereas just 10 years ago we were talking about Megabytes (10^{6}) stored on floppy disks.

Facebook, for example, stores photos. This statement is not impressive ... until you realize that Facebook has more users than the population of China. Facebook stored around 250 billion images in 2016. Can you imagine it? No, but really, try to imagine what 250 billion photos represent. So, in the big data area, when we are talking about volume, we're talking about astronomical amounts of data.

The volume of the data is enormous and a very large contributor to the ever-expanding digital universe is the Internet of Things (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 is created currently: Every minute we upload many videos on YouTube. In addition, billions of emails are sent, and billions of photos are viewed and upload, thousands of tweets are sent and billions of queries on Google are performed.

Remaining on the example of Facebook, 250 billion images are already doing a lot, but you have not seen anything yet: Facebook users have transferred more than 900 million photos per day, during the same year. Yes! By day, but this number looks actually like a drop of water.

This speed refers to the *velocity* at which the data is created, analyzed and stored. That appeal to the importance of the immediacy and instantaneousness to receive or transmit data from each of us and for all activities everyday compel businesses to improve their velocity reaction and anticipation. Especially, if we know that data comes in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. There are many different types of data and each type requires a different and specific type of analyses. Whatever the format (text, images, photos, videos …) a *variety* of data types needs storage, mining and analyzing.

Understanding big data means dealing with data volume that are significantly higher than those previously analyzed, at an incomparable speed, all while integrating a widely richer data variety (Sedkaoui, 2018a). So, in order to enrich the reader's understanding, we will take, in the following, each V with more details.

WHAT LIES BEHIND DATA VOLUME

The increasing computerization of all types of processing implies an exponential multiplication of the volume of data which counts now in Petabytes, Exabytes, Zettabytes, and Yottabytes. These units don't evoke you anything? Yet, it's thanks to them that the quantity of data produced is measured. A zettabyte, for example, refers to 10^{21}, so it refers to 1 000 000 000 000 000 000 000 bytes. That makes a lot? It is just a fraction of the amount of digital data produced around the globe every year.

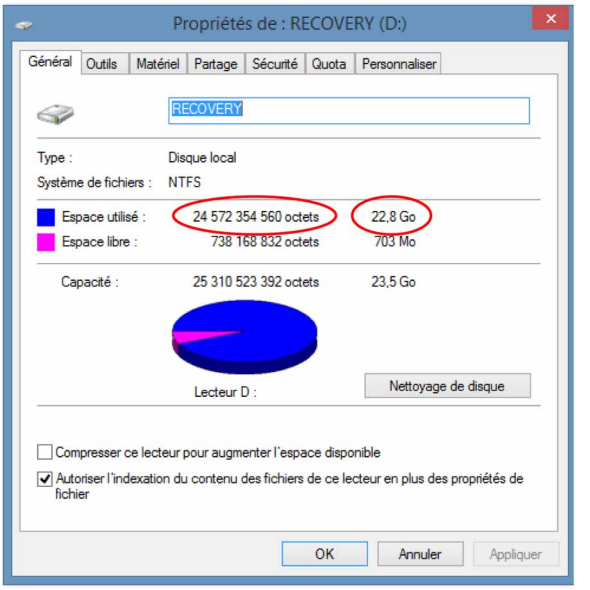
This volume is regarded as the first dimension for set points from big data. Let's discover how the data volume or the data size is calculated? And how big petabytes, exabyte, and zettabyte are?

At school (that was a long time ago), we all learned that: a kilobyte (KB) = 1024 bytes, and a megabyte (MB) = 1024 kilobytes, and so on. Then, a few years before 2000, a global organization found that it was too complicated, so decided that a kilobyte (KB) = 1000 bytes.

Here is an example; with the properties of my hard drive displayed (see Figure 3) from a Windows system (it's still not too old), to show how the data size is calculated.

Look at the values displayed; they are measured in bytes and in a gigabyte (GB), we notice immediately that the correspondence between byte and GB doesn't correspond to a simple scale.

Figure 3. Properties of a hard disk: Display in GB and in bytes



From a simple calculation, let's check what I said above:

- 24.572.354.560 bytes / 1024 = 23.996.440 kilobytes (KB)
- 23.996.440 KB / 1024 = 23.434 Megabytes (MB)
- 23.434 MB / 1024 = 22.8 Gigabytes (GB)

It is okay, we have calculated the values displayed or the size of my hard drive, what else now? What we mean here, is that companies must not rely on the size of their data, because the data volume is of little importance. And what really matters is how to make the most out of big data and generate value.

The several books about big data indicate, generally, the amount of data available around the world. You've probably read about it. For example, more data have been created in 2011 than in the whole history of Humanity, the number of photos held by Facebook is greater than the number of pixels ever processed by Kodak, and every day produces more videos than the first fifty years of TV, and so on.

But, the most important is not to carry some statistics about big data. It is a reality; the available data volume is huge. Results show that, in 2012, the world handled more than 2.8 zettabytes (2.8 billion gigabytes, an amount difficult to grasp).

This is more than what we have ever known, and the trend is increasing. But for companies that have to collect, manage and analyze the data available today, the crucial question is not about its quantity or volume. To take a snapshot, in a different context, "size does not matter". The data volume should not be the focal point. The important thing remains their analysis, that is to say, their transformation into knowledge, innovations, and value: the fourth 'V' of big data. In another way, the main challenge of using data is not in the collect, but in the choice of which data should be sought and how to make sense of it.

The big data principle is revolutionary and brings transformation possibilities to any business, but the term itself is problematic for a variety of reasons.

First of all, the notion of size, which refers to the word "big", is only one of the distinctive aspects of the new forms of data, and for most companies, this is not the most important characteristic, because the term raises several other questions. For many companies, solving the problem of lack of data structure proves more important than taking into account their volume.

The notion of size is obviously relative - what is great today will not necessarily be so tomorrow and what is great for one company is small for another, etc. If a database needs to be established, say big is one-tenth petabyte or more. Anyway, if there is one aspect of the data where "size matters", it will be related to the technology of data storage and process.

Big data has been defined according to the three mentioned Vs, and goes beyond these three Vs, by integrating other additional dimensions (see Table 1 in chapter 2). By the way, each of these additional dimensions starts with V. However, this definition is also problematic. Certainly, these dimensions are all important, but what happens if, for example, we only have one or two Vs? Does it mean that we cannot take advantage of big data?

These several interrelated Vs measures if the data has usefulness and is value relevant for its intended purpose and is also related to the quality of data (this point is more detailed in chapter 2). This is necessary because this can support decision making in an effective and efficient way, and data only matters when it is useful (Frizzo-Barker et al., 2016).

THE VARIOUS TYPES OF BIG DATA

Data is something that provides information about a particular thing and can be used for analysis. Nowadays, data can have different sizes but

different formats too. We cannot talk about big data without mentioning its different types. Big data is more than just a volume issue contrary to what its name suggests. In other words, we cannot understand big data without understanding data variety related to nature and formats. We are talking here about two notions (i) the variety of data types and; (ii) the structure of data.

In big data age, we have a variety of data form generated from different sources, and that data is more or less structured, and we will see that the variety of data is related to the data structure level. The big data age brings us to new types and formats of data and can be divided into the following categories (see appendix 1):

Structured Data

The term structured data refers to the data that is identifiable because it is organized or structured. Structured data concerns all data which can be stored in database SQL in a table with rows and columns. They have the relational key and can be easily mapped into pre-designed fields. It refers to the data that are stored in relational databases or corporate data warehouses. In another word, structured data are data which we established the functional sense and life cycle (creation of rules, values and possible technical means of representation). Also, this type of data is relatively simple to enter, store, query, and analyze.

Semi-Structured Data

This comes from the speed of arrival of the data, their volume, making structuring impossible. The messages in your mailboxes are a good example of semi-structured data. Semi-structured data is information that doesn't reside in a relational database but that does have some organizational properties that make it easier to analyze. So, this kind of data contains a structured part and an unstructured part, so it combines the two types (structured and unstructured). And due to unorganized information, the semi-structured is difficult to retrieve, analyze and store as compared to structured data. It requires software framework like Apache Hadoop to perform all this.

Unstructured Data

This is the simplest abstract form, a series of bytes. Unstructured data is also a description of a reality but whose codification, meaning, is not directly exploitable by the machine, for example, audio or video file, a text contained in a document or an email. Belong to this category, the data from social networks (Twitter and Facebook). Such type of data becomes difficult and requires advanced tools and software to generate value. Today, there is a predominance of unstructured data (they represent, according to the big data survey conducted by NewVantage Partner, more than 85% of digital data), which must be tried to organize a minimum (via metadata for example) to give them meaning.

Metadata

Describe and enrich unstructured data. In other words, the absence of structure is mitigated by the metadata. For example, the title and the tags to describe the videos on Dailymotion.

These new types of data may be intended to enrich the types existing before. And from it derive information and then producing knowledge, or which is called the target paradigm of "knowledge discovery", described as a "knowledge pyramid" where data lays at the base (see Ackoff, 1989).

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 (Sedkaoui, 2018a).

When analyzed optimally, each type of data brings valuable insights that business leaders can use to make accurate and timely decisions. Business leaders admit also that "the role of digital technologies is rapidly shifting, from being a driver of marginal efficiency to an enabler of fundamental innovation and disruption" (World Economic Forum, 2016).

Big data plays a very important role in the companies that value it. It comes to solve the problems of the slowness of processing, and the storage is much more robust and allows a mass storage of internal and external data.

The added value of big data is that it allows identifying the useful data and converts it into valuable information by identifying patterns, exploiting new algorithms, tools and new project solutions. The idea behind the term 'big data' is the one that justifies that we 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's the fact that these 3Vs change completely the way in which data is addressed, by putting it at the center of this transformation. The cases of many companies' experiences, such as Netflix, Google, Uber, Facebook, Amazon, Nike, and others (more detailed in chapter 3), illustrate that data can deliver value in almost any area of business.

WHERE DO BIG DATA COME FROM?

A big data approach makes it possible to enrich the data of the company with those of external sources. It is not a question here of tending towards diversity, but rather of giving oneself new perspectives on the business activity, the conjuncture in its sector, or its positioning on the Web. So, let's talk now about the variety of sources. Where do big data come from?

The answer is clear, they come from everywhere! This is also the reason that we observe today such a passion for the big data analytics. Mainly, there are two types of data:

- Private data accessible in a company, such as business databases, scanned documents, log files, data generated by IoT objects, etc.
- Public data accessible to everyone such as open data, social networking APIs and web scraping "extracting website content via a script or a program".

So, globally, data come from machines, humans and relational databases. With regard to machines, sensors of any kind will generate data without interruption, connected objects too, especially with the development of the IoT (Gartner, 2017). For example, in the field of computing infrastructure, sensors for gathering information from interconnected objects represent a new source of data. Their increasingly common use offers many opportunities and requires significant IT infrastructure needs (sensors, storage …).

Big data are also generated by humans. We can think about documents, images, presentations, videos, music, sound files or clicks flow corresponding to the navigation on the webs, Emails, and also the social networks which are a huge source of data. In addition, big data can be also presented as classic data relational databases.

Figure 4. Number of the connected object in the world

For example, cars all become connected to their environment and they send back data continuously. Multiple applications have emerged: for example, if we get the information on the triggering wipers of a large number of cars, we would be able to map in real-time the position and movement of thunderstorms.

Three examples of strategies to access this big data:

- Google wants to enter our cars and offers "Android auto", which, in the same way as with Chrome or Android, will allow the company to capture our data.
- Tesla offers a connected and increasingly autonomous car that integrates its own operating system. Tesla understands that it must keep control of the data produced by these cars.
- Offer an innovative mobility service based on a platform that collects all mobility information. This is the strategy adopted by companies like Uber or Lyft. Uber already offers this anonymized data, Uber Movement, to improve urban planning.

Every minute over the world, more than 700 people use an Uber, 18 million text messages, 481.000 tweets sent, 1.1 million profiles are swiped on Tinder and 266.000 hours watched on Netflix, 4.3 million videos viewed on YouTube, and more.

So, data is everywhere. We produce data in our daily personal and professional activities: we consume energy; we use applications to measure our sports performance; our cars transmit data, just like planes, trains, trucks; an e-commerce website manages its data; the industry becomes 4.0 to access the data Governance ... And we produce more and more data via sensors, cameras, radars, robots, drones and so on.

Globally, big data relies on four data sources:

The "Logs" From Traffic on the Company's Official Website

The company certainly has a showcase on the Web through its official website, which generates data traffic that must be analyzed. For a finer approach, and therefore richer in information, we will have trackers on the different pages to measure the navigation paths or the time spent on each page ... Other interesting questions, and therefore other sources of data, are the paths taken by visitors to reach the site: search engines, directories, rebounds from other sites ...

"Insights" From Social Media

Defining a digital identity and animating a community are now well-established practices. It is a source of data, competing with traditional questionnaire surveys.

Be careful however through the 'vanity measures', very easy to obtain (like, share, and retweet ...). Negative signals are less numerous, but express a strong gesture on the part of their author. So, remember to measure hidden posts and react! A complementary approach, combining quantitative and qualitative methods, is to collect comments from publications and to apply sentiment analysis algorithms to them.

"Third Party Data" or Behavioral Data to Better Target

Specialized Web players help you collect information about your customers or prospects and improve communication campaigns. Third party data is collected by these companies via forms or cookies. Beyond the classic identity information (sex, age, address ...), it is now much more effective to measure behaviors (navigation, hardware configuration, time spent on pages ...).

"Open Data"

Lacks of completeness, insufficient level of detail, relative seniority are the current flaws of many datasets. Open data is a field of investigation that should not be neglected; if only because of its low cost (the time spent searching!), and its inevitable development. 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.

So, data is everywhere, and become too complex and dynamic to be able to collect, store, analyze and manage with traditional analysis methods. 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 (Sedkaoui and Gottinger, 2017).

BIG DATA APPLICATIONS

Big data is it just a simple buzzword? Not at all! Our lives are already concerned in all their aspects by the uses of big data. Big data is already in several fields that the author lists as examples. Companies want to understand the behaviors and expectations of their customers to better target their proposals. They create predictive models to anticipate the departure of a client or sales of a product.

Understanding and optimizing processes apply to several areas: inventory management, human resources, optimization of delivery routes ... Big data applications also serve the individual. Online dating sites help to find a soulmate, connected clothes monitor our health and lifestyle.

Whether it's decoding DNA fragments or protecting premature babies, big data also has many applications in health, medical or pharmaceutical research. They help improve the performance of researchers, scientists or athletes. Security, fraud detection, optimization of traffic in cities, acceleration of financial exchanges, whatever one may think, the big data applications are there! It is a question of making an overview of the different applications in order to realize the transversal effect of big data. So, let's discover how big data is transforming different sectors?

Industry

The industry is one of the first sectors transformed by the emergence of big data. Predictive maintenance, enabled by the data revolution, makes it possible to anticipate the risks of breakdowns and defections of certain parts or components.

Data also makes it possible to optimize the expenses: one can optimize the energy consumption of a point of sale, optimize the costs of transport, stocks notably by crossing several types of data.

Telecom

The telecommunications sector is undergoing transformation thanks to Big Data. Some phone operators use big data and cognitive technologies to index thousands of documents and other images in just a few minutes. Thus, call center employees can solve problems much more quickly and efficiently. This strategy also enables businesses to realize significant savings. For every second earned on a call center call, companies in the industry save about 1$, or nearly one million dollars per year. Big data can also create personalized offers and packages based on historical data.

Manufacturing

In the manufacturing sector, big data can be used to avoid excess inventory and to create better products while saving money. Data analysis also makes it possible to build customized products for each customer based on their preferences and other individual criteria.

Retail

In the retail field, businesses can use big data to personalize the customer's in-store experience and provide a better quality of service. Sensors make it possible to determine the profile of each customer who enters the shop, the rays he visits, and the time he spends in each aisle. Similarly, the data of subscribers on the various social networks allow to improve the marketing, the design of the product, and thus to increase the incomes.

Sport and Fitness

Big data makes it possible to create smartphone applications for coaching for example. These applications include collecting data on users' training routines, exercises, calories burned, and much more. With this data, the apps can then coach users to help them achieve their goals.

Insurance and Bank

The finance sector is not left behind. It was one of the first sectors to benefit from big data, particularly for the issue of portfolio diversification,

by combining a multitude of real-time data to estimate the risks of each of the financial securities, in order to minimize the overall risk of a portfolio through diversification.

The advances in big data are of great benefit to the banking sector. The prediction of fraud is thus transformed by the arrival of the 3 Vs phenomenon. By combining multiple data on a borrower or an insured, with other external data, one can predict with great acuity the risk of fraud. In the same idea, big data can predict with a renewed precision what is called the risk of 'churn'. It involves identifying customer profiles that have a high risk of termination or attrition. We can then identify the sources of dissatisfaction and launch a loyalty campaign to recover these clients, once they have been targeted.

These developments have greatly improved the automation of processes; particularly with regard to claims management processes or loan agreements. The actuarial profession is also transformed by big data, the pricing of insurance contracts being done through a massive data set consisting of information on the competition, economic context, and risk assessment. In the insurance sector, companies are also using big data to reduce the time needed to process claims. Instead of the two days required so far, the treatment time can be reduced to just 10 minutes.

Tourism

The tourism sector uses big data to improve the customer experience. Airlines collect data about their passengers, such as allergies, food preferences or favorite location on the plane, and use each passenger's travel history to provide a completely personalized service. Some companies set up rewards programs for clients who share their data. Companies can also use predictive analytics techniques to determine when and where customers will travel to optimize pricing and service.

Government

For governments, big data makes it possible to improve the services provided to citizens. Citizens can easily access hundreds of thousands of documents allowing them to simplify administrative procedures. In addition, analytical technologies enable utilities to detect fraud attempts, analyze financial markets, conduct research for health and protect the environment more effectively.

Health

In the health sector, big data can extract relevant information from the medical history of each patient. Hospital admissions, doctor visits or emergency room visits can be analyzed to identify patients at risk of chronic disease, for example, to improve treatment and reduce readmissions. To take a concrete example, the use of a multitude of data can lead a country to anticipate the arrival of an epidemic. Indeed, thanks to the data collected on keywords, typed on google, for example, related to the symptoms of a particular disease and located in a defined geographical area, it is possible to know in real-time if a country or another is touch. This makes it possible to act quickly and efficiently and thus to prevent the spread of the disease to the greatest number.

Transport

Data is on the rise. The interest in its collection and analysis is growing because it's one of the key levers for the optimization and management of the transport sector. Early on, the sector has already seized new technologies to produce and analyze data. Embedded computing (geolocation, route planning, alert management ...), Transport Management System (TMS) or even the development of mobility tools (smartphones, tablets ...) have paved the way for big data. Big data is used to:

- **Understand:** Congestion phenomena and ways to avoid them are established through the analysis of user paths coupled to available alternative circuits.
- **Planning:** The definition of transport offers and infrastructure projects comes from the analysis of a mass of information concerning the data of the journeys: daily flow according to the day, the schedule, the route.
- **Know:** The profiles, needs of users are identified through the analysis of customer data via loyalty programs etc.
- **Optimize:** The shortest / fastest / cheapest route is found at time *t* based on real-time user data and data analysis.

Energy

This industry uses big data to deliver energy efficiency and reduce costs. By analyzing historical demand, it is possible to predict energy demand in real-time, thus effectively serving customers.

Agriculture

Big data is used by farmers to reduce losses, predict crop costs, the health status of livestock, or the number of pesticides needed. For example, giving animals nothing more or less than the amount of food they need can reduce costs and keep animals healthy.

Science and Research

Science and research are currently being transformed by the new possibilities big data brings. In physics, for example, processing billions of gigabytes of data at the LHC has led to one of the greatest discoveries in physics: the Higgs boson. Another example, the data collected around the world using satellites allows the design of meteorological models to better anticipate extreme events such as cyclones. Also, IT computing powers can be leveraged to transform so many other areas of science and research.

The data processed allows companies to obtain strategic advantages by taking into account a greater number of data, improving the existing (optimization and cost reduction, productivity gains ...), creating new products more targeted, or to improve the customer experience. Indeed, a better knowledge of new needs is clearly an asset.

For the technology sector, the creation of values inherent in big data seems obvious. Big data technologies are also used in the aerospace, automotive, chemicals, electronics, entertainment, hospitality, information, high-tech and oil sector, robotics, or logistics … Industrial maintenance, customizing offers, energy efficiency, preventive medicine or the autonomous car and more, are examples of the various applications of big data.

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 at more than $ 100 billion by 2030, over the world.

The profound transformations engendered by large-scale data processing capacity are able to redefine the boundaries between different sectors and companies that will be able to seize the opportunity. These varieties of big data applications do not stop there, and some uses are specific to each sector.

CONCLUSION

It is now observed that data or what is known as big data nowadays, is an important stake of business and entrepreneurial innovation and in particular thanks to artificial intelligence and machine learning. However, they are also a source of anxiety and distrust, the issues of which deserve our attention!

On the Web, we often see "big brother is watching you", but what we could understand by: "we spy on you", in reference to George Orwell's book (1984).

I read this book a long time ago, and I loved it. In this book, written in 1949, Orwell imagined a world in 1984, in which the inhabitants were constantly watched. Well, I think we can say since the advent of big data, we are constantly monitored, fortunately not for the same reasons as in the Orwell book! I hope you understand better now the illustration of this chapter: big data is watching you!

Big data is, therefore, a big challenge for companies. Like Orwell, we would have needed a whole novel to explain in words the sprawling world of big data. We have therefore turned to images (different figures illustrated in this chapter and throughout the book) because, thanks to a well-thought-out image, we can approach the many facets of the most complex subjects.

REFERENCES

Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, *15*, 3–9.

Chen, M. S., & Liu, Y. (2014). Big Data: A survey. *Mobile Networks and Applications*, *19*(2), 171–209. doi:10.100711036-013-0489-0

Frizzo-Barker, J., Chow-White, P. A., Mozafari, M., & Ha, D. (2016). An empirical study of the rise of big data in business scholarship. *International Journal of Information Management*, *36*(3), 403–413. doi:10.1016/j.ijinfomgt.2016.01.006

Gartner. (2017). Retrieved from: https://www.gartner.com/newsroom/id/3568917

McKinsey Global Institute. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Sedkaoui, S. (2018a). *Data analytics and big data*. London: ISTE. doi:10.1002/9781119528043

Sedkaoui, S. (2018b). Statistical and Computational Needs for Big Data Challenges. In A. Al Mazari (Ed.), *Big Data Analytics in HIV/AIDS Research* (pp. 21–53). Hershey, PA: IGI Global. doi:10.4018/978-1-5225-3203-3.ch002

Sedkaoui, S., & Gottinger, H. W. (2017). The Internet, Data Analytics and Big Data. In Internet Economics: Models, Mechanisms and Management (pp. 144-166). Bentham Science Publishers.

Sedkaoui, S., & Monino, J. L. (2016). *Big data, Open Data and Data Development*. New York: ISTE-Wiley.

World Economic Forum. (2016). *Global Information Technology Report*. Author.

Zikopoulos, P., & Eaton, C. (2011). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. New York: McGraw-Hill Osborne Media.

KEY TERMS AND DEFINITIONS

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.

Analytics: As 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 mounds of internally generated and externally available data.

Big Data: A generic term that designates the massive volume of data that is generated by the increasing 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."

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

Data Analysis: This is a class of statistical methods that make 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.

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

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

Internet of Things (IoT): The inter-networking 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: It 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.

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

Small and Medium Enterprises (SMEs): Are companies that fall under specific legal limitations regarding the number of employees and the annual turnover. However, this differs from one country to another.

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

Yottabytes: A unit of data storage, equal to one sextillion (102^4) bytes.

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

APPENDIX

Figure 5. Data types

Chapter 2
What the 3Vs Acronym Didn't Put Into Perspective?

ABSTRACT

Data sizes have been growing exponentially within many companies. Facing this size of data—meta tagged piecemeal, produced in real-time, and arrives in continuous streams from multiple sources—analyzing the data to spot patterns and extract useful information is harder still. This includes the ever-changing landscape of data and their associated characteristics, evolving data analysis paradigms, challenges of computational infrastructure, data quality, complexity, and protection in addition to the data sharing and access, and—crucially—our ability to integrate data sets and their analysis toward an improved understanding. In this context, this second chapter will cover the issues and challenges that are hiding behind the 3Vs phenomenon. It gives a platform to complete the first chapter and proceed to different big data issues and challenges and how to tackle them in the dynamic processes.

INTRODUCTION

Never trust anything that can think for itself if you cannot see where it keeps its brain.

J.K. Rowling. (Harry Potter and the Chamber of Secrets)

DOI: 10.4018/978-1-5225-7609-9.ch002

Big data defines the use of technologies and methods to analyze the different and voluminous available data. It is about identifying and making exploitable certain market trends, consumer behavior and so on. It is particularly used by business professionals to refine their targeting and analyze all facets of their consumer behavior. The data resulting from their purchases on the internet or in store, their preferences on social networks and their browsing history on the internet (cookies) thus serve as references to apprehend a global behavior.

Who collects data about customers knows more than the one that does not do that. It is logical. Yet, the benefit is far more important than 'knowing more' only. Data collection and analysis allows the business to update models, and to see trends, problems, and possible solutions. 'The result?' More loyal customers and a competitive advantage over the long term.

Going to the data culture to improve continuously and explore new use cases, especially under the impetus of the machine learning and artificial intelligence is a reality. But, the immaterial nature of the data formats and the volumes involved make big data a phenomenon that feeds fantasy and maintains public confusion and mistrust.

Data volume, data variety, and data velocity are the three criteria commonly used to define the big data phenomenon. It is the famous "3Vs rule". This short description suggests that the challenge around the data is only technological. However, it is not. The techniques used to analyze big data are just an extension of business analytics and business intelligence (see chapter four), two disciplines that have existing businesses for a long time.

What lies behind the expression big data is the general awareness, begun in 2012, of the strategic importance of data and the major changes they will bring to companies. Until now, data management was at the service of the business: it was a simple support function, which managers used to enlighten their lantern and guide the strategic choices. Today, things are reversing: the data becomes a strategic resource. The result is an upheaval in our business models, in which the data function and its extraordinary possibilities will drive business activity and create value.

As real gold mine, big data allowing companies to improve their processes, identify the needs of their customers and even anticipate their future consumption. However, to take advantage of this gold mine and exploit it properly, companies will have to ensure compliance with a roadmap, and be attentive to the major issues surrounding the notion of big data.

To engage in big data it is still necessary to take into account the data quality. Given the vast amount of information available, the relevant data must be identifiable and cleansed. Indeed, the databases contain their batch of bad data, these incomplete data or likely to be misinterpreted, which must be rectified. Also, a data quality audit should be the first priority of any big data project. In this respect, the technological solutions are multiple. Automatic correction tools exist for example and make it possible to ensure the relevance of the information collected and analyzed.

Also, one of the current challenges of big data is the development of complex tools to process and better visualize, analyze and protect huge data flows. These data come in bulk and come in various formats. The company must, therefore, invest in data integration solutions and the implementation of scalable infrastructures. Storage, in innovative tools (cloud computing, etc.) is in this respect preponderant. It must be coupled with software that uses sophisticated algorithms that allow the analysis of these large volumes of digital data in real-time.

In addition, most of the data collected by companies to define their strategy come from the private domain. Coming straight from the user accounts, this information affects the relationship between the company and its customers. The question of security around this data is therefore crucial because it engages the responsibility and reputation of the company.

The European regulation on data protection in Europe, the so-called "the General Data Protection Regulation (GDPR)", also provides for greater transparency on the use of personal data collected. New obligations are imposed on operators: they will have the obligation to ensure the consent of individuals (and to be able to prove it) for the collection and processing of their personal data. They will also need to implement the necessary security measures to prevent illegal processing or accidental loss of such data. In the age of social networks and connected objects, the processes of data protection are at the center of the debate between different issues.

The challenge, therefore, lies in the ability to extract value from the amount volume of data produced in real-time continuous streams with multiple forms and from multiple sources. All these challenges and others will be discussed in this chapter in order to clarify what lies exactly behind the 3Vs rule and the huge and diverse big data opportunities menu.

BIG DATA IS MUCH MORE THAN THE 3 VS

In chapter one, we talked about the phenomenon of big data to which we have historically associated the three primitives' Vs. The volume: for the amount of data produced and shared (collection, storage, and processing). The variety for the diversity of formats (text, audio, image, video), sources (sites, social networks, Radio Frequency Identification (RFID), Global Positioning System (GPS) etc.), and origins (structured, semi-structured, unstructured and internal and external data). And the velocity: of data collection, processing and sharing (in real-time).

You see? It all started with these 3 Vs. Three words intended to summarize the problems posed by the management information and demonstrate the inability of analysis tools to deal with such volumes, as varied and at high speed.

But, the question, that which allows understanding the novelty of this phenomenon is: what is really hiding behind this iceberg of data? Why do we talk about big data when we have always had data! This is true; we always had data, except that the difference, today, data come in large volume. For example, before, when a customer bought a book from a bookstore, the processed data was mainly related to the book (its reference, price …) and the purchase transaction (the total amount, the date of purchase …). Today, it is possible to buy the same book in electronic format from an online store. There are now vast data that revolves around this apparently simple purchase transaction. Here are some examples:

- The customer profile.
- The purchase history.
- All details of the navigation path leading to the purchase on the website.
- Comments and ratings made about the book on the retailer's website, Facebook, Twitter, blogs, forums, etc.
- The satisfaction survey data.
- The characteristics of the mobile device (tablet or smartphone) on which the book is viewed.
- The statistics of use of the book (frequency and speed of reading, parts of the book most engaging, least read, etc.)
- Geolocation data or where the book is downloaded, opened, etc.

Data quantity management is not a new problem. The 1880 US census required seven years of work to publish the results. In the census of 1890,

it would have taken more than a decade based on traditional methods. Only the invention by Herman Hollerith of a tabulator, ancestor of the computer operating with punch cards, helped to remedy the problem.

With the development of the Internet, digital and connected objects, this problem is back: data processing capabilities are currently lagging behind data collection capabilities.

For the specialists, the problem can be summed up by the three Vs rule:

The first problem that related to the large volume of data is the easiest to grasp. Every minute, a hundred hours of video is uploaded on YouTube. The majorities are never seen by and are of very limited interest. So, how to deal with this mass?

This failure to observe all the information has led practitioners to change the paradigm. Big data has changed the way we use data. Before the analysts had a priori idea about what they were looking for and verified it with the data, now they hope to extract knowledge without imagining it in advance. Until now, the models described the data, but, nowadays, the data generate the models.

The second problem is related to the variety of data which raises more technical questions: how can a machine analyze text, images, and video? In a group of people, for example, everyone can describe the same video in a completely different way. In this context, how make machines to have the same analysis approach? To this end, we need databases which can be recognized by all and can help to ensure the value of the knowledge extraction.

In the case of the Stock Exchange, for example, the codes associated with companies make it possible to identify them, even if several companies have the same name. This database facilitates the ultra-rapid exchanges that boost the stock exchanges.

This brings to the third problem that refers to the velocity and its economic stakes. Indeed, a search engine that would take ten minutes to display a result would become as practical as an encyclopedia in eight volumes. In order to be able to process queries efficiently, businesses activities need reactive structures.

Before big data, several companies were not able to capture, store or analyze all the data they would be interested in. Probably because of the unavailability of technologies that allowed it to be done at a reasonable cost and within a reasonable time. But now that the technological tools that have made the success of companies like Google, Yahoo, and Amazon become accessible, these companies will no longer see data the same way.

According to IBM (2012), the human species produces 2.5 quintillions (25 billion) bits of data each day by eating, sleeping, playing, working, and so on. This is the equivalent of two piles of DVDs placed on each other whose height goes from the earth to the moon!

So, knowing that these DVDs contain all the texts we send, the pics we take through the data of our every contact with all interconnected sensors and machines, this explains why data have become as 'big'.

So, when people talk about big data, they usually think that we're going to take a lot of that data, analyze it, and come up with something interesting. In fact, it's much more than that, because it also involves and consists:

- To take large amounts of data from different sources;
- To use these data of very different natures produced according to different rhythms without necessarily needing to 'translate' them into specific formats;
- To store this data so that it can be used at the same time for a whole lot of different analyzes corresponding to different objectives;
- And to do all this very quickly, and sometimes even in real-time.

Big data goes beyond the three known Vs. By the way, dear reader, I give you a few more Vs (Table 1).

The definition by the V is so classic that experts seek to explain every aspect of big data by additional V: Veracity (cleansing of noise in the data),

Table 1. The additional Vs of big data

Vs	Characteristics
Variability	A 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 to the analysis based on it.
Validity	How accurate and correct the data is for its intended use?
Vulnerability	The fact that a growing number of people are becoming switched on to the fact that their personal data is being gobbled up by the gigabyte, used to pry into their behavior and, ultimately, sell them things. (Marr, 2016)
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 endgame after addressing the other Vs. It refers to the benefit of big data that can be gained through appropriate analysis.

Source: Sedkaoui, (2018a)

Value (the analysis of the data must be motivated by the value business that it brings), Viability (looking for the 5% of the data that carries information), Variability (presence of inconsistencies in the data) and many more. Some researchers integrate seven additional dimensions, as shown in Table 1.

The additional V's all demonstrating other challenges involved in extracting value from unconventional datasets. Veracity, Variability, and Volatility refer to the problem that big data often consists of data that's accuracy is not assured, there are irregularities within the dataset, the data flow is uneven and navigating between the components is not straightforward. Veracity, which is often described as data quality and governance, is an important aspect of big data, as data not only comes from everywhere but belongs to everyone. Visualizations are needed to make sense of the findings. The driver of the whole process is to create Value, while the Validity of the findings or the Viability of utilizing them might be questionable (Shafer, 2017).

After clarifying the challenges related to the different Vs that lie behind the context of big data, now moving to those related to the quality, the complexity and the security of the data that crystallize the attention of businesses. First, let's talking about the data quality.

DATA QUALITY: SMART DATA VS. BAD DATA

The possibilities of the collection process of more date are now abundant, thanks to the multiplication of sensors, digitization, and connectivity. While it is now possible to store large volumes of data, another the challenge is related to its quality including a target of reliability and robust design. If the big data challenges are many (technology, budget, skills ...), the main suspect seems to be the data itself. Reports studies show that quality is the first barrier to deter decision-makers from integrating data into their decision-making process.

Between "buzzword" and reality, big data is more than ever a problem that drives companies. At the heart of these concerns, managing the quality of data is one of the keys successes when working with data. Data is the indispensable raw material for one of the new century's most important activities. Be careful! In your analysis and predictions procedure, a lot of data is not yet "the right data". There is, therefore, underlying difficulty behind big data, since more data doesn't mean necessarily better data.

Historically, since the creation of the first databases in the 1960s, companies had to deal exclusively with structured data, from a single and known source, while now they have to deal with a multiplicity of types coming from more

and more varied channels: computers, mobiles, tablets, GPS ... and eventually any connected object. These data comes mainly from:

- **The Web:** Newspapers, social networks, e-commerce, indexing, document, photo, and video storage, linked data, etc.;
- **The Internet and Connected Objects:** Sensor networks and Smart grid (cars, oil pipes, windmill turbines …), call logs;
- **Text Data:** Email, news, Facebook feeds, documents, etc.
- **Location:** GPS and mobile phone as well as Wi-Fi connection makes time and location information a growing source of data.
- **Science:** Genomics, astronomy, sub-atomic physics;
- **Commercial Data:** Such as transaction histories in a supermarket chain;
- **Personal Data:** Medical files;
- **Social Network Data:** Social network sites like Facebook, LinkedIn, Instagram …
- Public (open) data.

In this context, the great difficulty lies in understanding the diversity of these data, while it is estimated that more than 85% of data are heterogeneous and unstructured. In another hand, the big challenge of big data is to make the right information available to the right person and cross it with other data sources. This requires an adaptation of technologies and organizations.

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 that data.

Data is critical for the BI and data mining trades. However, the professionals who handle them are not immune from a bad appreciation of their quality. In this context, we can cite two cases in which this error could occur:

- First of all, a change in internal processes can be a source of damage for data management. For example, a simple change in the cleaning method or a data processing system during different operations increases the risk of degradation. This could also include the arrival of new staff, less qualified, the change of technique or simple tools.
- In the second case, it is considered that the introduction of new data, whether the updating of already existing information or the introduction of new data, could be a source of quality error. The explosion of

phenomena such as social networks will only increase data traffic. Publishers will even propose solutions to analyze them and increase the ROI.

However, can we really say that large-scale data processing would be profitable for our businesses? Have we thought about possible future issues?

With regard to this last question, companies are also challenged on the usefulness of big data. Referring to the term "Garbage In, Garbage Out" (GIGO) it is believed that collecting and analyzing bad data will produce distorted results that can be disastrous for the company. Once the implementation of the first data warehouse in the 1990s the question of the quality of the data was a major issue. In the US, the theorem GIGO was immediately widespread (Sedkaoui & Gottinger, 2017).

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.

Big data processing requires an investment in computing architecture to store, manage, analyze, and visualize an enormous amount of data. It is indispensable for many activities. But it is important to be prudent in our analysis and predictions because a lot of data is not yet 'the right data'. 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 'Big' term data to 'Smart data' term since the effectiveness of companies' strategies now depends on the quality of data (Sedkaoui, 2018a).

So, the quality of data is a problem that persists. This topic had been answered very much upstream of the operating chain in the reprocessing of the data entry interfaces. Today, the influx of data and immaturity around the new big data technologies leads us too often to ignore in the processing, the data considered incomplete or erroneous. But are these 'irrelevant' data at first glance really? Are not they just as valuable and untapped because of their unreliability?

For many companies, we must get rid of 'unusable' data: missing data, addresses not updated, unknown product purchased from the catalog, false email address … But this unstructured data can also be useful sources of information for your business project.

From my point of view, it's not a question of 'bad data' it's only about the 'data quality' as exploiting these bad data which themselves contain exploitable information. The solution is to make these bad data more reliable by eliminating duplicates, by straightening those that are poorly formed or incomplete to bring them to the same level of quality as the others. Mixed with other data, they then form a whole which may apply the business requests of their users.

It is then important to be able to qualify the quality level of the data in a flow feeding the data platform of a company to identify the bad data, to apply them to the ad hoc reliability processing, and to integrate them in order to have a more important source of data to analyze.

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, processing and the way data is managed. It's the "analysis" that will be applied to data which will justify big data, not the collection of data itself.

WHAT ABOUT DATA COMPLEXITY

It seems unthinkable today to talk about big data without mentioning the four major factors that converge in the highly innovative information technology sector and commonly known as SMAC, for Social, Mobile, Analytics, and Cloud. Some predict that these factors will combine with new cognitive and language technologies, also known as "machine learning" to have a major impact on the economy. In other words, the big data is directly powered by four generating forces:

- **The Web and Social Networks:** With the new paradigm of the Web and user generation of data on the Web, embodied by social networks;
- **The Mobile:** With the smartphone, every human being generates an increasing amount of data: location, email, photos, videos, tweets …;
- **The IoT and Sensors:** Connected objects are another source of raw data, which retrieves a large amount of data through their sensors. The Internet of things 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 & IDC, 2014). 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. Connected devices can not only connect to the Internet, they can also connect and share information with each other. In fact, machine-to-machine (M2M) connections will grow to 27 billion by 2024 (Marr, 2017).

- Finally, we've already talked about it; it's *the Open Data movement*: public and quasi-public administrations are required to make freely available to all the actors all the data they generate and process.

With these four generating forces, there is an avalanche of data generated every moment by humans and machines. As the information technology of a new generation based on IoT, cloud computing, and mobile internet, big data realizes the record and collection of all data produced in the whole life cycle of the existence and evolutionary process of things. Although the term has been in use since a decade it is only with the rise of web 2.0, mobile computing and the internet of things that organizations find themselves increasingly faced with a new scale and complexity of data.

In the early 1980s, while the subject of the 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 (British unit of measurement which worth approximately 2.54 cm). In 2000, it is anticipated that this capacity will be "1.25 x 1011" bytes per square inch. The vision of Hal B. Becker has largely concretized since 2011, data was recorded Zettabytes, 200 times more in a single year than what had been measured previously.

Most parts of the data available today are unstructured. In addition, it comes in different formats due to the variety of data sources; dealing with the data formats variety, increases the *complexity* of storing and analyzing.

Also, the data we handling is usually represented by a series of characters and numbers. For example, the text you are reading now is coded based on the numbers and letters. The sequence of the decimals of π is represented by a series of digits. The nucleotide chains that make up DNA (support for the genetic information of living) organisms) are encoded by a series of letters {A, C, G, T}; thus, TGTCCCCAATTA describes a DNA sequence. The series 10010100101001100101 describes twenty binary draws, for example, 0 for stack and 1 for face, etc.

However, in the machine, all these series are recoded into a sequence of bits (binary symbols {0, 1}). To evaluate the raw information content of a series of characters, we need to know the number of bits needed for its

encoding. For example, the above DNA sequence can be encoded by coding each character on 2 bits: A by 00, C by 01, G by 10 and T by 11. Its raw content in information is 24 bits.

It's to say, that, due to the IT advent, modern datasets are evolving not only in term of size or volume or even diversity but also in term of 'complexity'.

In addition to the Garter's and McKinsey's big data definition mentioned before, EMC Education Services (2015) argues that there are three attributes standing out as defining big data characteristics:

Huge Volume of data: Rather than thousands or millions of rows, Big Data can be billions of rows and millions of columns.

Complexity of data types and structures: Big data reflects the variety of new data sources, formats and structures, including digital traces being left on the web and other digital repositories for subsequent analysis.

Speed of new data creation and growth: Big data can describe high-velocity data, with rapid data ingestion and near real-time analysis.

Complexity measures the degree of interconnectedness and interdependence in big data structures such that a small change in one or a few elements can yield very large changes or a small change that ripple across or cascade through the system and substantially affect its behavior, or no change at all (Katal et al, 2013). Traditional software tools are not enough for managing the increasing volumes of data. Data analysis, organization, retrieval and modeling are also challenges due to scalability and complexity of data that needs to be analyzed. Cukier and Mayer-Schoenberger (2013a) proposed the transformation of three thoughts in the big data era:

1. It is not random samples but the whole data;
2. It is not accuracy but complexity;
3. It is not causality but correlativity.

'The whole data' refers to the transformation from local to overall thought, taking all data (big data) as analysis objects. 'Complexity' means to accept the complexity and inaccuracy of data. The transformation from causality to correlativity emphasizes more on correlation to make data itself reveal the rules. It is closely related to the understanding of things by complex

scientific thinking, which is also the integral thinking, relational thinking, and dynamic thinking.

The volume and complexity of big data require in turn a change in computing paradigm, both in how we structure data and how we process it. Managing and storing such large amounts of data requires advanced technologies: event processing platforms, sophisticated database management systems, and specialized analytics algorithms. In relation to the definition, the reason why there is intense complexity in big data lies in the amount of data that is traveling across the internet today along with there are also: ambiguity, viscosity, and virality related to the 3Vs.

- **Ambiguity (Uncertainty):** Emerges when there is less or no metadata in big data. An example can be a graph or something that usually needs a description. For example in the large volume of data M and F can be taken for March and February instead of male and female.
- **Viscosity (Consistency):** This term is often used to describe the latency time in the data relative to the event of being described. For example, social media monitoring falls into this category, where a number of enterprises just cannot understand how it impacts their business and resists the usage of the data until it is too late in many cases.
- **Virality:** Describes how quickly data is shared throughout a network among people who are connected. The measurement result is the rate of spread of data in time. For example re-tweets on a tweet.

So, 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, smart devices as well as social networking, data in a company is more complex because it includes not only structured traditional relational data, but also semi-structured 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 comes 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, this data shows a great variety, which often reflects the complexity of the studied phenomenon.

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

SECURITY AND PROTECTION ISSUE

Globally, when we talk about big data, we first think about volume, variety, and velocity commonly referred to as '3Vs'. These three characteristics show that the big data is a mixture of pure 'storage' and 'analyzing' of data, and therefore requires the implementation of measures to guarantee the security of data in terms of integrity, conservation, classification and access control. So, for each of these three characteristics, security remains a difficult goal to achieve.

While big data regularly become an important issue of research and has been used everywhere in many industries, big data security has been increasingly concerned. Security is one of the main criteria for the success of a big data project. Without it, disastrous incidents can impact businesses and their activities.

A growing number of businesses are using big data technology to store and analyze petabytes or even zettabytes of data, including weblogs, navigation path data, and social media content, to better understand their customers and their businesses activities. As a result, the classification of information becomes even more critical; and the ownership of the information must be determined to allow for an acceptable classification. Nevertheless, there is a noticeable challenge between big data security and the uses of big data:

- This is a new technology for companies, and any technology that is not well controlled introduces new vulnerabilities.
- Typically, big data implementations use the open source code, which means that backdoors and default credentials may not be recognized.
- The attack surface of the nodes in a cluster may not have been examined and the servers adequately reinforced.
- User authentication and access to data from multiple locations may not be adequately controlled.
- Legal obligations may not be respected, with problematic access to log files and control tracks.
- The risks of introducing malicious data and inadequate validation of data are important.

The security problems related to big data are multifaceted regarding the origin of the data, the loyalty of their collection, the objective pursued (common scientific good or competitive advantage), the transparency or the opacity of

the goals pursued, the storage infrastructures and implemented calculations and the open or closed character of algorithmic processing.

Big data is not new and many experts are looking at the issue, including the US Federal Trade Alliance or the renowned analyst firm Gartner. For the latter, in 2018, half of the companies will experience data theft due to poor security practices. Achieving security issue in big data has, therefore, become one of the most important barriers that could slow down the spread of technology; without adequate security guarantees, big data will not achieve the required level of trust (Thuraisingham, 2015), so the 3Vs phenomenon brings big responsibility (Rijmenam, 2014).

The vast majority of data comes from the many devices and machines reporting to each other and to those running them. From the assembly line at the manufacturing plant to the passenger jet in flight, millions of bytes of data are generated and then analyzed. Some of the captured data is 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.

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 (Sedkaoui & Gottinger, 2017). Thousands of data shared by individuals expose their privacy more than ever before. Personal data is worth gold for marketers, financial institutions, employers or governments. The consequences can be disastrous for individuals, who may, for example, be denied a job or credit because of their data.

The Cloud Security Alliance (CSA) (2013), which is dedicated to promoting best practices in cloud security, has already detailed the key security challenges. According to CSA, there are, principally, four different aspects of big data security (see Table 2):

Table 2. Big data security: Main challenges based on CSA report

An Aspect of Big Data Security			
Infrastructure Security	**Data Privacy**	**Data Management**	**Integrity and Reactive Security**
Secure Computations in Distributed Programming Frameworks	Privacy-Preserving Data Mining and Analytics	Secure Data Storage and Transaction Logs	End-point validation and filtering
Security Best Practices for Non-Relational Data Stores	Cryptographically Enforced Data Centric-Security	Granular Audits	Real-time Security Monitoring
	Granular Access Control	Data Provenance	

- Infrastructure security,
- Data privacy,
- Data management, and
- Integrity and reactive security

This division of big data security into four principal topics has also been used by the International Organization for Standardization in order to create a security standard for security in big data. It continues the reflection with a new report (CSA, 2016) that proposes the ten best practices to meet these challenges and thus helps to inform the market about the security of big data.

Among the tips provided, there is the creation of a trusted environment in Apache Hadoop-type distributed programming frameworks through Kerberos authentication, or equivalent. This network authentication protocol relies on a mechanism of secret keys and tickets, without resorting to passwords that are too often intercepted. Another advice from the CSA related to the fact of hiding or delete all identifiable data (names, addresses, social security numbers, among others) to ensure perfect anonymization of data. The report also alerts to the presence of additional identifiers such as postal code, date of birth, or gender, which may also identify the nature of the data.

Also, beware of data warehouses; often poorly secured, they should be combined with strong encryption solutions, with separate storage space. The authors of the report also believe that companies using non-relational data warehouses such as NoSQL databases are at a disadvantage because these products generally do not have advanced security features.

They also advise the use of real-time security and compliance monitoring solutions, privacy analysis systems, verification of data provenance, use of cryptographic techniques, and more.

Big data: 'opportunity' or 'challenge' for data security and privacy? That is the question today. Between those who see 'big hacking' or 'big brother' and those who consider it as a 'big opportunity', big data is a promising subject in both cases according to the use that is made of it.

The advance in big data analytics brought us tools to extract and correlate this 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 of that where linking the provided data to the IMDB movie reviews allowed to identify some users.

Data security has been widely investigated over the past thirty years. However, today we face new issues in securing and protecting data, which

result in new challenging research directions. According to Bertino and Ferrari (2018), some of those challenges arise from increasing privacy concerns with respect to the use of such huge amount of data, and from the need of reconciling privacy with the use of data. Other challenges arise because the deployments of new data collection and processing devices, such as those used in IoT systems, increase the attack potential.

Today's threat environment imposes the three Vs of big data. Each of these is increasing at an astounding rate and has required a shift in how security vendors manage threats.

The diversity of data sources, formats, and data flow, combined with the streaming nature of data acquisition and high volume create unique security risks. So, security issues are magnified by the 3 Vs of big data. The use of large-scale cloud infrastructure with the 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 additional other Vs of big data amplify not only analytics tools but also security management challenges.

Data security not only involves the encryption of the data but also ensures that appropriate policies are enforced for data sharing. Security aims to preserve digital systems against malicious actions to affect confidentiality, integrity or availability of the system itself (Sedkaoui, 2018a). An effective cybersecurity 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 the condition of confidence, digital security is a major challenge for companies, that the advent of big data, cloud, and mobility in an environment increasingly connected makes it even harder.

The needs of new security challenges because of the growth of security intelligence require big data and big data analytics.

BEYOND DATA SIZE ANOTHER ELEMENT BECOME IMPORTANT: "THE TIME"

Big data is not just a buzzword. The data analysis ecosystem has been entirely built to be able to evolve with the latest technological advances and to easily

handle huge amounts of data. Big data is created digitally and collected automatically. Berman (2013) identified six different mechanisms through which data can be founded:

1. The data is already collected in the course of normal activities and is waiting to be used. The data owner does not want to discover or to do anything new, but to do better what it has always been doing.
2. The data is 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 like Amazon.
4. A group of entities that have large data resources federates 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 exist before big data.

Every second, many visitors interact with companies' websites leaving behind a tremendous amount of information that companies can then use to create customized experiences. Faced with such a challenge, it's important to make sure that the technologies they use are able to process this volume of data correctly! To describe the current economic climate, we can say that it is data-centric. Companies use data from a variety of sources: social media, news, financial markets, economic trends, system usage statistics and a slew of connected 'objects'. The problem with all this data is that companies do not know how to use them or do not know if they are stored and used on an ad hoc basis.

But, what happens if companies, for example, want to offer a live promotion via the mobile device, depending on the profile and location of their customer? Or if they want to redirect drivers according to the traffic problems caused by the weather? In another way, how to meet the challenges related to the real-time?

So, the challenge is no longer to collect the data, to exploit, to process and to analyze it better, in such a way that allows businesses to generate knowledge in order to upgrade the process of decision making and achieving higher performance. Beyond the advent of ICT and of increased data production,

Table 3. Big data real-time processing

Family	Input Data	Results	Latency	Implementation Example
Batch	Files, results of a query (HDFS, Sqoop ...)	The results will only be available at the end of the processing,	Often of the order of a minute.	MapReduce
Micro-batch	Small files, Web API, ...	results will only be available at the end of micro-batch processing,	Often of the order of the second.	Spark streaming
Real-time	Small files, Web API, ...	results are available as and when	Sometimes less than the second.	Flink, Tez, Storm

dissemination, and processing speeds, another element has recently become critically important: "time".

As the size of the data sets to be processed increases, it will take more time to analyze. In some situations, results of the analysis are required immediately. So 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 (Sedkaoui & Gottinger, 2017). In the big data age, there are three big families of processing (see Table 3):

- **Batch:** The process will analyze all the available data at a time t.
- **Micro-Batch:** Analyze all available data every n seconds.
- **Real-Time:** Data processing as they become available. It is not necessary to wait for the end of the input data to issue a result.

Batch processing makes it possible to analyze a set of input data until the source is exhausted. As long as data are available, the process will continue and we will have a coherent and accessible result only at the end of the treatments. In order to avoid this tunnel effect, it is possible to split the input data and this is where the 'incremental' notion is important. It will allow the new data to be taken into account without the need to reprocess all the data already processed. A perfect example is Map Reduce in its Hadoop version.

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 (Sedkaoui, 2018a). Then, 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.

The importance of time carries with it a notion of information circulation speed (Sedkaoui & Monino, 2016). The notion of real-time is quite relative and depends on the context: milliseconds, seconds or minutes. Batch and Micro-batch solutions are distinguished from real-time solutions. In the case of Micro-batch analysis, for example, results are produced every *n* second. For real-time analysis, each entry is processed immediately and produces a result. Apache Spark is a micro-batch solution while Storm and Flink are streaming solutions.

Overall, holders of real-time offers all sell more or less the same thing, more or less packaged way, and the main message is the same: collection, storage, processing and visualization of data, all in real-time!

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 push the door of the analytics in real-time.

Traditional data processing architectures know how to collect data before transforming it and then analyzing it. These three operations are globally performed one after another, a waltz 3 times in sum! Traditional approaches cannot be used for real-time data analysis because they rely on historical data only. Today, companies are strongly interested in using external sources of data, including social media, blogs, and sensors. This external data is essential for many activities that give companies advantages.

Big data is gaining momentum to become "fast data". While many companies only consider the contribution of data processing in volume, others associate this vision with a real-time component. Big data real-time analysis may seem simple at first glance; it's about managing data streams. But a lot of fresh data can also enrich historical information. This complicates the process of data analysis. We must manage both instantaneousness and the past data, by limiting the memory footprint, and not necessarily maintaining the completeness of the data.

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

This evolution could not fall 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 on the Web of more and more businesses and individuals but also by the rapid development of the IoT. This digital world doubles in size every two years: by 2020, predicts IDC, it will reach 44 zettabytes (ZB) (44 billion GB!).

In descriptive, predictive, and prescriptive analytics (more detailed in chapter four), 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 objects to monitor medical symptoms with information in medical records, analysis tools can alert health professionals if proactive action is urgent for the patient.

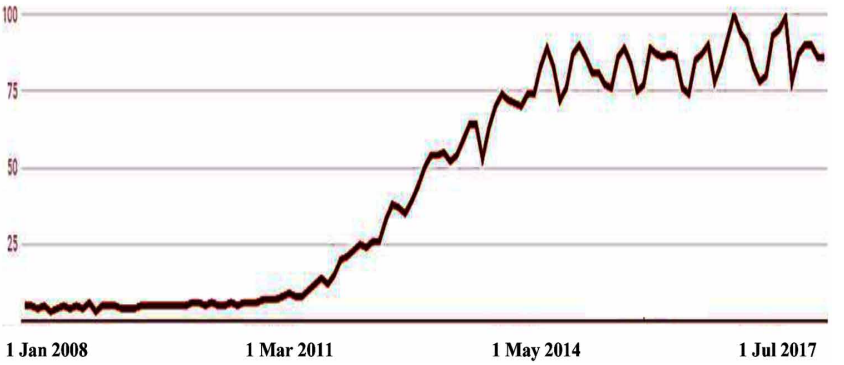
However, some analysts and specialists are beginning to wonder. They do not question the potential of real-time analytics but question whether real-time is still necessary or useful. Most of the knowledge obtained today comes from data stored and analyzed later. This is partly because there is more data, but also because it allows for quality assurance through proper data preparation. Thus, the data can be used at the most opportune moment.

The question isn't so much whether real-time data analysis is possible, but whether real-time may be the "right time" for that specific data and knowledge. Data analysis must support decision-making. The analysis at the right time, therefore, provides guidance to the right person as soon as it is needed. Sometimes this is done in real-time, but often it is not. In other cases, a combination of stored and real-time data is needed to generate useful information, for example, on the evolution of customer behavior over the time.

Adopting real-time analytics is therefore not always the best answer for accelerating data analysis and knowledge acquisition. Instead, one can focus on improving data preparation and management capabilities. This aspect often represents the longest and most difficult part of the analytics process. In other words, an enriched data management platform helps companies to improve their analytical capabilities, as it promotes better use of existing data.

But - because there is a 'but' - I think this process leaves out a fundamental point: data, like food, have an expiry date. The longer they reside in your data store, the less up-to-date they are. And, therefore, they are potentially less useful. In the health sector, for example, information about a particular patient should be recent. Data older than one year will not be used because the patient's condition may have deteriorated significantly or improved since. But data covering the past month or the six past months, showing a trend, and are certainly more useful.

Figure 1. Big data search term relative popularity on Google



In practice, spending time trying to find the 'perfect' use of data can be counterproductive. Paraphrasing Nike, we could say "Just do it!" That is to say, start experimenting with some data, and recognize that the failure is good, since it allows you to test other options. Manipulating models, trying to perfect them, is actually the worst way to quickly get information from data. Perfection is often the enemy of 'good enough', which is especially true with data and modeling.

MAPPING OUT BIG DATA ISSUES AND CHALLENGES

Data is flowing ever faster, almost in real-time. Moreover, their format has never been so varied: digital data but also texts, images or videos to mention a few. Big data is often characterized by the three Vs (sometimes four or even seven or more) (Wedel and Kannan, 2016; Marr, 2016, Sedkaoui, 2018a): volume, velocity, and variety. The first two Vs are important from the IT point of view (storage and computing), while the latter is important from an analytical point of view.

On the other hand, some argue that big data is just a fad that will go away like the others. Analyzing the popularity of the search term big data on Google, we discover that it has grown exponentially since 2008, and has stabilized since 2015. Marr (2016) says that the trend around big data and his name could disappear, but the phenomenon will continue and will only gain speed. According to him, the data will simply become the "new standard" within a few years, when all organizations will use them to improve their business and how to conduct it. We would not have said it better.

However, understanding and acting on a growing volume and variety of data is not easy. Wedel and Kanan (2016) express it more formally and argue that companies have invested far too much in data collection and storage, but not enough in their analysis. Although big data is a top priority for many companies, few of them are currently gaining value. Thus, in this book, our intention doesn't only focus on 'big data', but also the aim is to insist on the analytical dimension: indeed, to profit from big data, it is necessary to put in sets a good strategy for data analysis.

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

Big Data Dimension

Many researchers have discussed and suggested various big data issues related to its three known dimensions, in the literature. Each dimension presents both challenges and opportunities for data management to advance decision-making. This 3 V's provide a challenge associated with working with big data. The volume put the accent on the storage, memory and computes capacity of a computing system and requires access to a computing cloud. It would take a very long time to transfer data from multiple data sources to the cloud and back from cloud to processing point. To overcome this transferring issue two methods have been proposed (Katal, 2013). First, just process the data in the same place where it is stored and only transfer the required information. More specifically, bring the processing code to the stored data instead of transferring the stored data to the processing code, known as MapReduce algorithm (Kambatla et al., 2014). Second, transfer only the part of the data which seems more critical for analysis (Katal, 2013). 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 that large variety of input data (Sedkaoui, 2017, Sedkaoui, 2018b). Existing infrastructure, machinery, and techniques are not capable to process such amount of data in real-time. Although some advanced indexing schemes (like FastBit) (Wu, 2005) and processing methods like MapReduce (Dittrich et al., 2012; Triguero et al., 2015) is available to boost the processing speed but processing of Zettabytes and even Exabytes of data is still a challenging task.

Technological Context

There are many challenges of using and implementing big data. Manyika et al. (2011) indicate that a key obstacle is the consistency of internal and external databases, implying that there is a 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, semi-structured and unstructured). The most successful companies understand the limitations of the technology behind their big data operation and recognize the importance of combining analysis with a sound understanding of the context, a good intuition for the industry, and a critical attitude towards insights derived from data.

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 (Sedkaoui, 2017). The major challenge to overcome is management's lack of understanding of the potential value big data can bring to companies (Morabito, 2015). The goal was to manage the increasing amount of data, information and to ensure its usage and flow across the companies. Data is required to be managed in different steps and most of all analyzed (Kudyba, 2014), for organizations to gain knowledge and value. Management of data includes tasks like cleaning, transforming, clarification, dimension reduction, validation etc. Companies can make the use of business intelligence to manage a large amount of data, for example, quantum computing and in-memory database management systems allow economically effective and quick management of large datasets (Buhl et al., 2013).

The challenges include not just the previous contexts, or other concerns described previously, but also other issues related to scalability and heterogeneity which represent also big concerns.

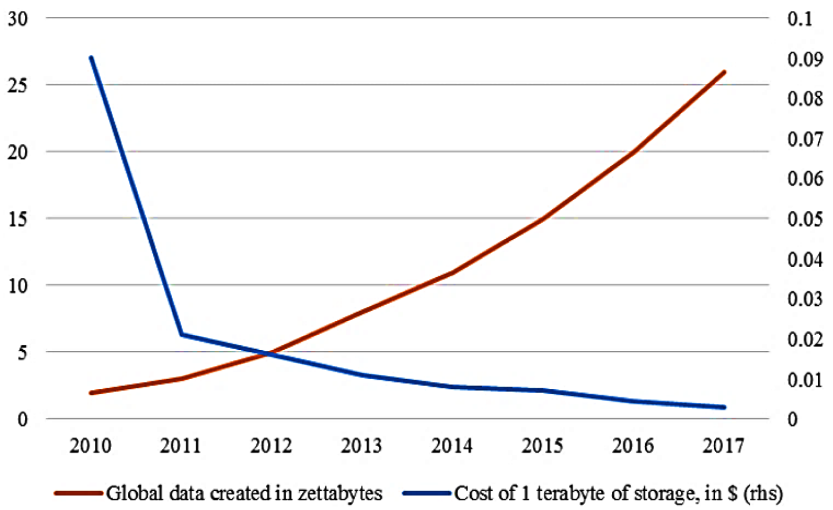
- **Heterogeneity:** Data can be both structured and unstructured. They are highly dynamic and do not have a particular format. It may exist in the form of email attachments, images, pdf documents, medical records, graphics, video, audio etc. and they cannot be stored in row/column

format as structured data. Transforming this data into a structured format for later analysis is a major challenge in big data analytics. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, 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 BD 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 (heterogeneous mixture data issue). Especially, in the case of complicated heterogeneous mixture data, the data has not only several patterns and rules but characteristically, the properties of the patterns vary greatly (Fujimaki & Morinaga, 2012).

The nature of existing data (greatest dimension, diverse types, the mass of data, structured and unstructured ...) 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

Figure 2. Costs of storage and data availability (2009-2017)
Source: Reinsel, Gantz and Rydning (2017); Klein (2017)



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), and analysis of data through the availability of targeted databases, software, and algorithms. Companies are required to deal with these several issues to be able to seize the full potential of big data. So, data in itself isn't a power; it is its using that gives power, and more one gives an exchange of data and information, more one receives (Martinet & Marti, 2001).

Another important element that needs more attention, is the one related to the gap between the data production and the capacity of analyzing that volume.

As a business, it's the consequences of this data explosion that you need to care about. In another word, the volume of data is not much important than your ability to do something with data and create value from it.

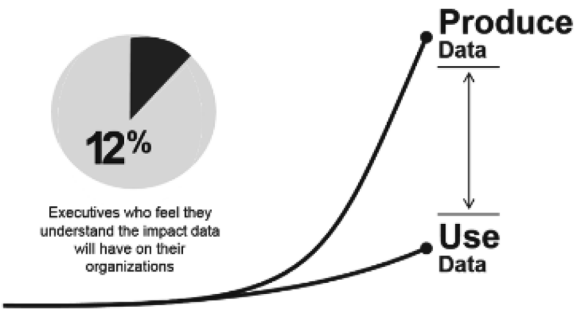
Many companies have not yet grasped how big data analytics can improve their performance (Johnson, 2012), a statement verified by Brown et al. (2011) who argues that companies need to address considerable challenges to be able to seize the potential with big data. Such statements show that there is a gap between the positive prospects of the large volume of data and the actual knowledge and use of big data in businesses today.

This appeals you to accord a great importance to two key results:

- **Knowledge Gap:** *The difference between collecting data and understanding data*

This idea can be seen in the knowledge pyramid which illustrates that data are essential to build the knowledge in order to make a good decision, and

Figure 3. The useful data gape
Source: https://practicalanalytics.files.wordpress.com/2011/12/usefuldatagap.png



generate value. In fact, the emergence of big data has a potential to influence a company. According to Frizzo-Barker et al., (2016), it has a potential to change the thinking of companies about data infrastructure, business intelligence and analytics and information strategy (Brynjolfsson and McAfee, 2011).

By using big data, companies are able to measure significant more about the business in their context and it is possible to make translations of that knowledge, which can improve the decision making and therewith the performance of the company (Brynjolfsson & McAfee, 2011). So, if you went to act like most companies, then you need to invest both time and energy in order to address the knowledge gap using analytics, business intelligence, and data warehousing.

- **Execution Gap:** *The difference between understanding data and acting on it*

To better understand big data and respond to the problem, we need to introduce the concept of 'the capacity to act". In another way, it is about bridging successfully the gap between the steps of collecting and understanding the large available data to the integration of efficient solution to better conduct business strategy.

Be able to act effectively on divers' aspects of data analysis techniques and IT tools give you the power to marrying big data analytics and business entrepreneurial insights to manage the huge business opportunities in your field.

But, it should be noticed that it is hard to bridge both these gaps at once, focusing on just the knowledge gap will cause 'knowledge impotence', i.e., too much knowledge and too little action. For example, you may have solved the knowledge gap by knowing the answer to a simple question like: How many customers do you have in total and how many have you lost? However, an execution gap may still exist, because, for example, you may be aware of a competitive threat but not have a good way to operationalize your response. This execution gap can kill the competitive advantage of your business activity.

CONCLUSION

If oil was the fuel for the industrial society, the data is the new oil needed to build the knowledge society. The proliferation of data, with the advent of social networks and connected objects, their centralization thanks to the web

and the cloud and the distribution of their products which becomes more and more collaborative and continuous, new challenges surface volume, variety, and velocity of data storage and processing. This involves significant constraints, both technological and business.

The advent of big data comes to answer these challenges that go beyond the traditional decision-making because the data are not only structured and do not come only from the information systems of the company, but especially from external varied sources.

As you will have understood, the analysis of big data becomes a major stake allowing new economic applications: to know the consumer perfectly in order to decrypt his behavior of consumption. The first challenge is then to improve the analysis tools to better interpret this ever-growing flow of data. Figures that we do not understand are not helpful. Nevertheless, the central problem is not the efficiency of data analysis. Indeed, the other challenge that big data must overcome, and certainly the most important, is the respect for privacy by precisely delineating personal data that cannot and should not be exploited by third parties. But rest assured the picture is not so black! The lucrative domains do not have a monopoly on big data, as shown by the humanitarian association Mobilizing Health, which uses them to provide care in India.

REFERENCES

Becker, H. B. (1986). Can users really absorb data at today's rates? Tomorrow's? *Data Communications*.

Berman, J. J. (2013). *Principles of big data: preparing, sharing, and analyzing complex information*. Amsterdam: Elsevier, Morgan Kaufmann.

Bertino, E., & Ferrari, E. (2018). Big Data Security and Privacy. In S. Flesca, S. Greco, E. Masciari, & D. Saccà (Eds.), *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years. Studies in Big Data* (Vol. 31). Cham: Springer. doi:10.1007/978-3-319-61893-7_25

Big Data Working Group, Cloud Security Alliance (CSA). (2013). *Expanded Top Ten Big Data Security and Privacy*. Available online: https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf

Big Data Working Group, Cloud Security Alliance (CSA). (2016). *Cloud Computing Top Threats in 2016*. Available online: https://downloads.cloudsecurityalliance.org/assets/research/top-threats/Treacherous-12_Cloud-Computing_Top-Threats.pdf

Brown, B., Manyika, J., Chui, M., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The Next Frontier for Innovation, Competition, and Productivity*. Washington, DC: McKinsey Global Institute.

Brynjolfsson, E., & McAfee, A. (2011). *Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy*. Lexington, MA: Digital Frontier Press.

Buhl, H. U., Röglinger, M., Moser, D. K. F., & Heidemann, J. (2013). Big data. *Business & Information Systems Engineering*, *5*(2), 65–69. doi:10.100712599-013-0249-5

Cukier, K., & Mayer-Schonberger, V. (2013a). *Big Data: A Revolution That Will Transform How We Live, Work and Think*. Boston, Ma: Houghton Mifflin Harcourt.

Dittrich, J., & Quiane-Ruiz, J. A. (2012). Efficient big data processing, Hadoop MapReduce. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, *5*(12).

EMC & IDC. (2014). *The digital universe of opportunities: Rich data and the increasing value of the Internet of Things*. Academic Press.

EMC Education Services. (2015). *Data Science & Big Data Analytics*. Indianapolis, IN: John Wiley & Sons.

Frizzo-Barker, J., Chow-White, P. A., Mozafari, M., & Ha, D. (2016). An empirical study of the rise of big data in business scholarship. *International Journal of Information Management*, *36*(3), 403–413. doi:10.1016/j.ijinfomgt.2016.01.006

Fujimaki, R., & Morinaga, S. (2012). The Most Advanced Data Mining of the Big Data Era. *Advanced Technologies to Support Big Data Processing, 7*(2).

IBM. (2012). Global Business Services, Business Analytics and Optimization Executive Report. *Analytics: The real-world use of big data*.

Johnson, J. (2012). Big data + Big analytics = big opportunity. *Financial Executive*, *28*(6), 50–53.

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing*, *74*(7), 2561–2573. doi:10.1016/j.jpdc.2014.01.003

Katal, A., Wazid, M., & Goudar, R. H. (2013). Big Data: Issues, Challenges, Tools and Good Practices. *IEEE Spectrum*, 404–409.

Klein, A. (2017, July). Hard Drive Cost Per Gigabyte. *Backblaze*.

Kudyba, S. (2014). Information Creation through Analytics. In S. Kudyba (Ed.), *Big Data, Mining, and Analytics. Components of Strategic Decision Making* (pp. 17–48). Boca Raton, FL: CRC Press Taylor and Francis Group. doi:10.1201/b16666-3

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The Next Frontier for Innovation, Competition, and Productivity*. Washington, DC: McKinsey Global Institute.

Marr, B. (2015). *Data Strategy: Beyond the big data buzz: how data is disrupting business in every industry in the world*. Academic Press.

Marr, B. (2016). *Big Data in Practice (Use Cases) - How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results*. Wiley. doi:10.1002/9781119278825

Martinet, B., & Marti, Y.-M. (2001). *L'intelligence économique: Comment donner de la valeur concurrentielle à l'information*. Paris: Editions d'Organisation.

Morabito, V. (2015). *Big data and analytics: strategic and Organizational impacts*. Springer International Publishing. doi:10.1007/978-3-319-10665-6

Reinsel, D., Gantz, J., & Rydning, J. (2017). *Data Age 2025: The Evolution of Data to Life-Critical*. IDC White Paper.

Rijmenam, V. (2014). *Think Bigger: Developing a Successful Big Data Strategy for Your Business*. New York: Amacom.

Sedkaoui, S. (2017). The Internet, Data Analytics and Big Data. In Internet Economics: Models, Mechanisms and Management (pp. 144-166). Bentham Science Publishers.

Sedkaoui, S. (2018a). *Data analytics and big data*. London: ISTE. doi:10.1002/9781119528043

Sedkaoui, S. (2018b). Statistical and Computational Needs for Big Data Challenges. In A. Al Mazari (Ed.), *Big Data Analytics in HIV/AIDS Research* (pp. 21–53). Hershey, PA: IGI Global. doi:10.4018/978-1-5225-3203-3.ch002

Sedkaoui, S., & Monino, J. L. (2016). *Big data, Open Data and Data Development*. New York: ISTE-Wiley.

Shafer, T. (2017). *The V's of Big Data and Data Science*. Elder Research Data Science and Predictive Analytics.

Thuraisingham, B. (2015). Big data security and privacy. *Proceedings of the 5th ACM* Conference on *Data and Application Security and Privacy*, 279–280. 10.1145/2699026.2699136

Triguero, I., Peralta, D., Bacardit, J., Garcia, S., & Herrera, F. (2015). MRPR: A MapReduce solution for prototype reduction in big data classification. *Neuro Computing, 150*, 331-345.

Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, *80*(6), 97–121. doi:10.1509/jm.15.0413

Wu, K. (2005). Fastbit: An effcient indexing technology for accelerating data-intensive science. *Journal of Physics: Conference Series*, 16.

KEY TERMS AND DEFINITIONS

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.

Analytics: 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 mounds of internally generated and externally available data.

Big Data: A generic term that designates the massive volume of data that is generated by the increasing 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): 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.

Cybersecurity: Also known as computer security or IT security, is the protection of computer systems from the theft or damage to the hardware, software or the information on them, as well as from disruption or misdirection of the services they provide.

Data Mining: This practice consists of extracting information from data as the objective of drawing knowledge from large quantities of data through automatic or semi-automatic 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.

Garbage In, Garbage Out (GIGO): In the field of computer science or information and communications technology refers to the fact that computers, since they operate by 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.

Internet of Things (IoT): The inter-networking 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.

Machine Learning: A method of designing a sequence of actions to solve a problem that optimizes automatically 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 includes, but is not limited to cellular communication, Wi-Fi, ZigBee, WiMAX, wireless LAN (WLAN), generic DSL (xDSL) and fiber to the x (FTTx).

Return on Investment (ROI): Is a performance measure, used to evaluate the efficiency of an investment or compare the efficiency of a number of different investments. ROI measures the amount of return on an investment, relative to the investment's cost. To calculate ROI, the benefit (or return) of an investment is divided by the cost of the investment. The result is expressed as a percentage or a ratio.

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. This 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.

Web 2.0: This term designates the set of techniques, functions, and uses of the world wide web that has 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.

Chapter 3
Big Data Applications in Business

ABSTRACT

Nothing seems to stop the big data revolution. At the same time a promise of a better world and anguish of a possible big brother, big data is the new reality of the digital economy: it is the new territory of development and creation of value for the companies. The opportunities seem endless, which is why we must appropriate the data to better understand and tame it, in order to prepare for the future towards which it seems to lead us. After the theory, let's go to the "fun" part with some examples of big data uses that you may know without realizing it. We will see in this chapter some examples of using big data in a dynamic improvement of the business strategy in order to generate value.

INTRODUCTION

Someone is sitting in the shade today because someone planted a tree a long time ago.

Warren Buffett

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

DOI: 10.4018/978-1-5225-7609-9.ch003

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

Beyond what is big data and how it may impact the business context, this third chapter highlights the big data applications within businesses that have not been discussed in the previous chapters. Through this chapter, we recall the importance of big data in the business context in conducting decision-making, and the role it plays as a complement to the creation of new opportunities for businesses.

In recent years, large companies have begun collecting and analyzing very large amounts of data. Today, they realize that big data can give them a competitive advantage. One of the advantages of this phenomenon is to be able to consult huge amounts of information very quickly. And that is good because, regardless of the industry, companies have a consistent need to know what customers really think in order to track trends.

If large volumes of data are used correctly, it becomes easier to access information such as purchasing behavior or consumer preferences. It is then possible to zoom in on data segments to study them more closely.

After definitions of the different concepts in the first two chapters of this section, here is the time for the application of big data in the business context. But, it should be noticed that, in the big data universe and when working with data, the level of maturity is very different depending on the business activity and the company. The Web companies started very early as well as those which were in competition with pure players. They go through several phases, ranging from information on technologies and their integration to generalization and implementation through the implementation of Proof of Concept (PoC), tooling and deployment for some business applications.

Major players like Google, IBM, Cisco or Microsoft have invested for several years in the construction of Datacenter but have also deployed solutions dedicated to data analysis. However, new entrants are looking to take a piece of this cake that is envious. In addition, many companies operating in more traditional sectors will be able to take advantage of the big data revolution.

WALMART: BIG DATA AT ALL RAYS

When it comes to data; and is phenomenal amounts; it's hard not to think about Walmart and its expertise in data management that has allowed it to

optimize its distribution processes to dominate by the costs in its field of activity. In fact, Walmart is the symbol of a traditional business; as opposed to a virtual company like Google or Facebook …; where data management is based on data analysis.

All Walmart's business decisions are based on the extraction of strategic information from data generated by the consumption habits of its customers and the products of its inventory.

No company better illustrates the advantages of leveraging massive volumes of data for competitive advantage than Walmart, which operates a data warehouse with, at last count, 583 terabytes of sales and inventory data built on a massively parallel 1,000 -processor system from data-warehouse-technology vendor Teradata, an NCR Corp. subsidiary. While some companies might consider having more than half a petabyte of data overkill, at Walmart it's the way to do business (Babcock, 2006).

This allows Walmart to:

- Store products or forecast the number of human resources required, based on an analysis of trends in the consumption habits of its customers;
- Use just-in-time management methods to manage inventory procurement in partnership with suppliers, helping to keep storage costs to a minimum;
- Know in real-time where a product is in the supply chain and how long it will take to find it on store shelves.

In fact, Walmart's data analysis is a strategic advantage that the company realized long before its competitors when it invested 4 trillion dollars in 1991 to:

- Create the '*RetailLink*' program, which allows it to share information with its suppliers this program has transformed Walmart's business model;
- Set up its database, which allows it to inventory and consults the performance of the products of all its branches in real-time and according to a multitude of criteria from which it can cross-analyze;
- Implement all kinds of innovations such as barcodes and EDI technology (Ludloff, 2011).

At the internal level, data analysis allows Walmart to optimize its operations. At the external level, data analysis allows the company to identify all kinds of correlations and trends to better understand clients and predict their needs.

In this respect, the best example dates back to the August 2004, Hurricane Charley hit Florida hard. The balance will be 34 people in Jamaica, Cuba and the United States and damage recorded 16 billion. A few weeks later, a tornado named Frances threatens the state again. *Linda M. Dillman*, director of information at Walmart (major chain of distribution in the United States), then presses here teams: no question of waiting, we must anticipate avoiding stock-outs of basic necessities. Teams will then analyze the recovered data just before Charley's move a few weeks earlier and they will make an astonishing discovery.

If, as one might expect, people massively bought bottles of water, candles, and batteries, they also realized that sales of Pop-Tarts had increased sevenfold just before the arrival of the tornado. It will not have been long before entire trucks arrive in the local Walmart loaded with precious sesame. Walmart has learned its lesson well as it is now known around the world for its massive use of big data to anticipate the needs of its customers.

It's not over; working with big data allows Walmart to be able to identify a strong correlation between purchases of baby diapers and beer by analyzing sales receipts (Sedkaoui & Monino, 2016). Walmart blew up sales of diapers and beer after identifying that between 5 pm and 7 pm, the shopping cart for young men in all of its stores often contained baby diapers and beer. The rearrangement of merchandising at the point of sale dramatically increased the sale of these two items. This real-time correlation system operating on a constant flow of data allowed Walmart to increase its sales by simply placing diapers and beer closer together on their shelves.

Walmart manages its technologies like a true high-tech company with its '*WalmartLabs'* located in Mountain View in California only a few blocks from Google headquarters, thanks to its innumerable innovations in self-scanning, mobile apps, and big data; Walmart has information on 145 million Americans; and its bulimia of buyout start-up ...

In 2004 the Bentonville group announced they have acquired "*Stylr"* mobile app that allows users to find clothing nearby. An app, developed in the "Silicon Alley", the New York competitor of "Silicon Valley" ... that began to have its small success in the USA among fashionistas.

The number one in the US, in this field, has begun to pace itself. Experiments are multiplying: buy-outs of start-ups, the creation of incubators or investments

in hardware and software ... Walmart has thus offered the company *Kosmix* to build its own infrastructure for the real-time study of data.

Walmart can be considered as a true pioneer in big data. Since 2011, Walmart has been developing many tools in his WalmartLabs and has more recently acquired a Data Cafe (Collaborative Analytics Facilities for Enterprise). With more than 2.5 petabytes of data analyzed each hour, Walmart has one of the largest data collections in the world and can anticipate the needs of its millions of weekly customers.

If big data is a key driver of the customer experience, it also makes it possible to optimize the management of the supply chain. And for this, the notion of real-time is paramount. For example, Walmart uses its cloud to track millions of transactions daily. The demand for each product, the inventory levels, the activity of the competitors is studied with a fine comb, in order to modify the prices to keep the most competitive rates. According to Walmart, the Data Cafe would have made it possible to pass the problem solving to 20 minutes on average, in 2016, against 2 to 3 weeks ago.

If Walmart is a forerunner, the big data revolution is affecting the entire retail world today. The exploitation of the data transforms the models in place and impacts both the supply chain and the knowledge and the customer relationship. For distributors, this is an opportunity to seize in order to differentiate their activities with an ever richer and innovative customer experience.

AMAZON: ARTIFICIAL INTELLIGENCE INTO YOUR EVERYDAY LIFE

More data is above all a more precise targeting and which takes into account a more objective behavior on the web than on a single site. If you connect to e-commerce site only for a specific product, it will only have a narrow view of your interests, so it will go to feed data from news sites, social networks, and others for better understand what interests you in terms of information and web searches to recommend articles to you live.

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 the machine learning application, 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.

Thanks to machine learning algorithms that anticipate the evolution of our needs from our previous visits to Amazon.

The real-time marketing works on this principle, with a need for 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.

Machine learning 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 linked 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 the learning phase and make Machine Learning accessible.

Latest: Amazon *Echo*. No need to move or go to the web to shop, a simple voice exchange with '*Alexa'*, virtual assistant, is enough.

At the base installed in the heart of the intelligent home assistant Echo, artificial intelligence *Alexa* tends to become unavoidable. Alexa is 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 it easier to use, such as starting or stopping a wash cycle, checking and adjusting refrigerator temperatures, starting a recording on its connected TV, etc.

Amazon has already launched "*Look*" or a small connected camera, able to analyze your look and tell you if you are well or badly dressed. Jeff Bezos's company goes further with an artificial intelligence program based on an algorithm able to design a garment. It's a sort of a fashion creative artificial intelligence in a way (Sedkaoui, 2018a).

In 2016, Amazon represents a larger market than most of the major players in the US market combined.

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 machine learning. Developers can use them to quickly create face, object, or activity identification, tracking, or content detection apps.

Amazon also developed concepts that everyone thought were still reserved for the fantasy. This is the case, for example, of the impressive cashless grocery stores that are part of the "*Amazon Go"*. Customers can simply use the

displays, and a complex system of chips, scales, cameras, AI, and transmitters and receivers coupled to smartphones customers can track, validation, and payment of items without the user noticing.

This results in a considerable saving of time, 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 the customer or by product, or the possible increase in expenses related to the simplicity of the concept and its fun aspect. The first store is opened in early 2017 in Seattle, for Amazon employees exclusively as a test phase.

This is impressive, how Amazon gets this success? It's obviously the data. Since its creation in 1994, the company has adopted a culture largely driven by data. Thanks to the knowledge given by the data, the customer gets the right product at the right time and is satisfied with the image of the logo of Amazon: a smile.

NIKE: FROM SPORTS EQUIPMENT'S SALE TO SEGMENTATION OF THE MARKET BY BIG DATA

As another example, in the sports equipment industry, Nike has long been regarded as a leader of the market. This company sells sports equipment, clothing, and accessories. However, the way this company acts, in recent years, has changed. This company is heavily invested in applications, connected objects, and big data.

Nike is known as the world leader in many categories of sports shoes, and more generally in the global sports equipment market. The firm is also renowned for its use of technology for design, manufacturing, marketing, and sales.

Nike provides its clients with a complete 'ecosystem' focused on health and well-being. The data sources used (see, for example, Nike+ and FuelBand) provide a personalized service for clients and allow them to monitor their physical activities. In exchange, the company gains access to real-time intimate data of their clients' habits and uses, which can be, in turn, used to improve its product offerings.

The brand offers 13 different ranges and its presence extends to 180 countries. However, it is the way it segments and serves these different markets that really differentiate it. Nike calls this strategy the 'category offense'. Rather

than divide the world geographically, the brand divides the world according to sports efforts. The theory is that people who play golf, for example, have more in common with each other than those who live close to each other.

So far, we can say that this philosophy worked well. Since Nike adopted this strategy in 2008, sales have increased by 70% (Marr, 2015). This marketing and retail technique are largely guided by big data. The firm has notably invested in data thanks to its connected objects. Although the firm stopped selling her FuelBand bracelet in 2014, the company is building partnerships with other wearable brands including Apple, which recently launched the Nike + Watch Apple.

From a prospective point of view, we can go further and imagine that tomorrow, with the development of 3D printing, even mass products, should be customizable without increasing the costs of scale. Nike, already at the initiative of the customization with Nike id, has for example invested in 3D printing to be able to offer customers tailor-made and customized shoes. We would then return to the era of mass-customization.

The Nike iD brand has launched a new concept which allows their clients to create their own shoes online. Their marketing strategy included several campaigns such as the iPhone application and Nike Photo iD. With Photo ID, clients can send a photo that the

Nike designers, may subsequently, create. Nike iD also launched a poster campaign in Time Square; customers could use their phone to customize a pair of Nike, they later received a message with a wallpaper of their creation and had the opportunity to order the shoes. This is an example of an innovative interaction of a brand with its customers, thanks to digital.

Nike has made a radical shift in its approach by connecting its products, building a data management infrastructure to support the processing becoming a service provider and not just a product. Despite the limited success of its connected watch, continues to work on the subject, including exploring the track of smart textiles.

Nike's big data projects do not stop there, because in 2015, as part of a conference dedicated to investors around the partnership between Nike and the NBA, CEO *Mark Parker* explained his willingness to connect sports fans to the action they see in the field and learn more from athletes in real-time. To do this, the firm deploys many efforts in the field of big data technologies.

Effectively, in order to promote their new LunarEpic running shoe, Nike has launched a unique campaign. The brand proposed to the inhabitants of Manila to run against their own avatar.

Nike has created a track in the shape of the new shoes. The runners' shoes were equipped with sensors, which allowed to recover the characteristics of their race for the first round. Then in the second round, they ran against their avatar displayed on a LED screen. This avatar actually represented their time of the previous turn. The runners were thus competing against their own silhouette. The data from the first round was retrieved through the sensors and sent to the screen via radio frequency technology. At the end of the race, coaches gave personalized training to runners.

This award-winning campaign also allowed Nike to develop its philosophy of surpassing consumers. The track had a climb, sharp turns ... It was not easy to tackle. This philosophy has also proved that Nike is always at the forefront of technology in its field.

A SUCCESS STORY OF NETFLIX

Netflix, founded in 1997, is the video-on-demand service straight from the United States. Formerly confined to broadcasting existing content, the company now produces its own series. Netflix's particularity? Analyze everything you look at and deduce trends for future productions. That's how the House of Cards series was born: the basic idea was to create a remake of a political mini-series (namesake) broadcast on the BBC in 1990. By analyzing the likes of subscribers who liked this first version, the teams of analysts of Netflix realized that subscribers had also watched a lot of movies starring Kevin Spacey or directed by David Fincher. This is where the project of this award-winning series came about.

How does it decide what content to provide, going as far as to make personalized offers? By making a progressively enriched analysis of our purchases and consumption habits that ultimately defines our tastes with a high degree of precision (Sedkaoui and Monino, 2016).

The Netflix model is largely based on a deep knowledge of its subscribers as a constant source of service improvements. Many still think that big data or "hyper-personalization of the customer experience" are only abstract concepts. The strength of Netflix, like Facebook or Google, lies in the personalization of its recommendations. This is what has allowed the market for film rental/series to reinvent itself.

A few years ago, you had to go to a DVD store (or distributor) (and VHS ...) to watch a movie. A personalized recommendation could take place in the next film only if you know the vendor. Netflix is proud of its recommendation

algorithms. The US online video platform communicates regularly on 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 seems to you quite familiar and that illustrates well the "Netflix addiction".

At Netflix, there are many examples of big data applications, and we will mention, in this chapter:

Application One: Test the Design

A/B testing applied to the design and ergonomics of a website is not a revolutionary technique. Nevertheless, at Netflix, the data and design teams work hand in hand. Any modification of the platform, whatever it is, was first tested by the users. Netflix ensures that each change will actually improve the customer experience.

Application Two: Guide the Creative Choices

One of Netflix's original series crystallizes fantasies about their uses of data. Indeed, the hit show House of Cards was entirely based on analytical data. Using big data, Netflix was able to create an algorithm that allowed the company to discover several things about users:

- Several had viewed The Social Network by David Fincher in full.
- The British version of House of Cards had good ratings.
- Those who listened to the UK version of House of Cards also listened to Kevin Spacey and/or David Fincher.

These three factors join a significant number of subscribers to the service offered by Netflix (Leonard, 2013). That's why the company decided to produce an American version of the series directed by David Fincher starring Kevin Spacey. This decision proved fruitful as House of Cards is one of the most lucrative series of the company.

It should be noted, however, that these analyzes are not self-sufficient (another mix of popular successes could have been quite far-fetched ... Intuition, talent, and creativity remain indispensable for the creation of an original series. Take for example another Netflix creation, the Unbreakable series Kimmy Schmidt, which is not the result of a statistical prediction but of the unfathomable spirit of Tina Fey.

Application Three: Recommend Content

The force of the war! The famous Netflix recommendation engine allows 75% of subscribers to select a movie or series. Their customization algorithms are therefore very effective, which may give the user the impression that the application knows them better than they know themselves. Again, we must recognize the limitations of the algorithm. If the data gathered are considerable, and from various sources (geo-location, detailed uses, device used, metadata ...), it is not an exact science. The subscriber remains unpredictable.

Given these few examples of big data application, it's easy to see how Netflix uses customer knowledge to personalize and optimize its service. The customer experience is therefore dramatically improved: subscribers spend less time searching for content, and more time watching. What makes hitch any amateur movie or series ... It comes back to binge-watching.

The algorithms used by Netflix coupled with machine learning (through the collection of data to evolve constantly) will allow creating diversity for the proposed visuals depending on what the user is likely to click on 'Reading'. Let's take a closer look at how data is 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 …) but also sociological specificities and dozens of other criteria.

So, the importance occurs when Netflix begins to consider other factors, such as when and how movies are viewed (device type, the 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 this data is collected, it remains only to cross and segment it to obtain groups of users (clients) with distinct characteristics. This is where this algorithm comes in. It will mix all of these data with the components specific to works in the Netflix catalog in order to offer almost certain movies and series to which user will be interested.

From analyzing the way a user is consuming and of his passage from one series to another or from one movie to another, the algorithm will present him a personalized 'path' inadequacy with his profile and observe his reaction.

Take the example of a user who has a preference for the genres "action movies" because of his readings. One of the operating rules of Netflix will

be to propose a movie of the same kind of 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 in a recommended movie.

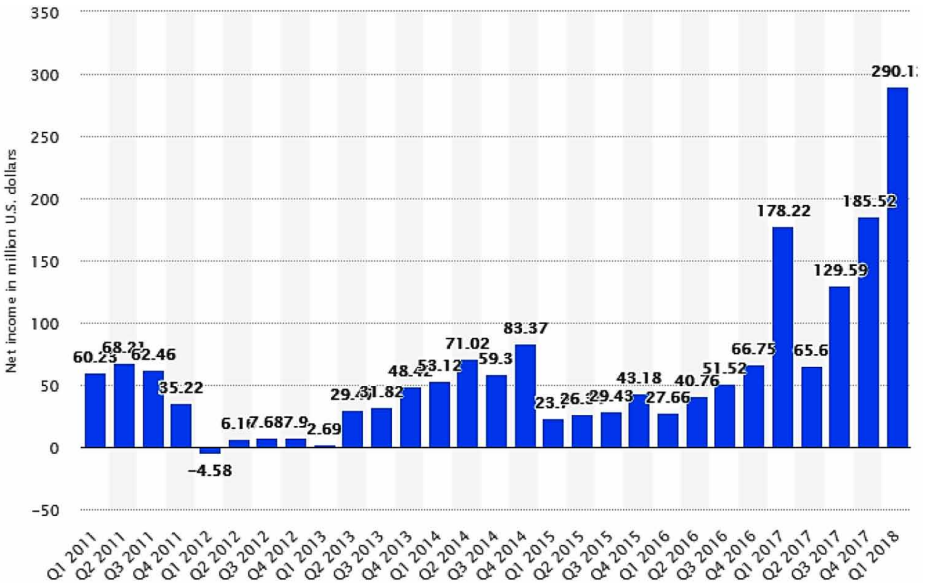
Thus, this work of recommendation and personalization by the illustrations proves a real challenge for the engineers of Netflix. Many of them work continuously by screening the millions of personal data available (Sedkaoui, 2018a).

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 dozens), as for technical teams, who face the challenge of managing 20 million requests per second with a short time latency.

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, Netflix continues its personalization work.

Figure 1. Netflix net income from Q1 2011 to Q1 2018 (in million $)
Source: Statista Netflix, May 2018



With more than 117 million subscribers in the world, Netflix net earnings increase in the first quarter of 2018, like it's illustrated in Figure 1, and the video rental/streaming company reported net earnings of 290.12 million U.S. dollars. Why do we like so much about Netflix? Let's say that big data and the resulting personalization encourage us to binge-watching!

WELCOME TO THE UBER TRACK

Many companies are increasingly using big data analytics to improve their business. Companies carry more and more data but do not always know what to do, much less how to make them valuable. With the emergence of big data emerging also tools capable of storing and building reports simply and dynamically. Companies that are born in the digital context understand these issues, especially when it comes to a company like Uber.

Uber is one other example of successful use of the latest analytics technologies. *"There are only four people/organizations in the world who know my location at all times"*, writes Ron Hirson, for Forbes: *"my wife (because I tell her), Apple (because of Siri), the NSA (because of NSA), and now Uber"*. According to him, the American VTC company is turning into a 'Big Data Company'.

The Uber smartphone app connects passengers and drivers with a principle based on big data analytics (crowdsourcing). Anyone who is willing to offer his driver services can offer his 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. By 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 Airlines Company and hotel chains to adjust the price on demand in real-time through predictive analysis.

To adjust to the demand that varies in time and space, the competitive advantage comes from the ability to motivate a workforce to be responsive in real-time. It must connect to the request. It is channeled by platforms. The

role is to meet this demand continuously according to a certain standard. For Uber, it is to deliver the service within five minutes.

The challenge is therefore to pilot a variable fleet of VTCs, to guide a population by giving them capabilities (to point out where the right opportunities are), by distilling incentives (overprices), and by monitoring their behavior (ratings).

From a managerial point of view, this system proves very effective: it generates more satisfaction at a lower price. It gets rid of the heaviness of the HR and management control, it outsources the production and its control but it does not lose the hand in terms of management. This system achieves it by subtle means of technology and persuasion.

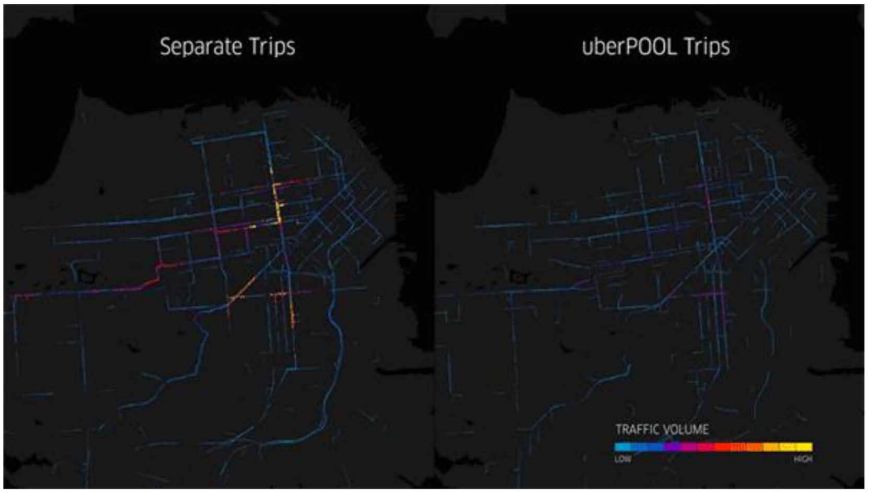
Knowing where there is the most passage, allows determining a map of value in the cities. This success has not escaped to the market leader of the VTC. The American company launched in 2017 'Uber Movement', a portal where cities can register and receive personalized statistics on the road traffic in their streets from the anonymous data collected by the smartphones of its drivers and users. Present in more than 450 cities, the app has so far convinced Boston, Manila, Sydney, Washington and more.

The big data strategy within Uber concerns also the data visualization. Uber now wants to optimize urban travel times in major cities often congested with traffic jams. Initially, the company was mainly interested in setting up a demand-based pricing system (see the Geoserve system based on the importance of supply and demand depending on the location). Then these connected data were used to optimize the circulation of drivers to finally lead to a new way to understand mobility.

Data visualization, in turn, helps the audience understand what Uber does and how it works. On the one hand, it helps to understand how Uber uses GPS data, and on the other hand, it concerns what you can do for the general public. Uber teaches us a lot about using large data. They make it possible to reveal the urban rhythms according to a real cartography of the 'pool' of the city according to the sectors and the schedules of circulation.

The giant of the VTC has opted for a multi-cloud strategy. Oriented micro-services, its architecture combines many development languages, databases, tools ... In less than ten years, Uber has become the most famous unicorn in the world. To provide the smoothest possible user experience, Uber has sought to reduce the friction of its application while simplifying it as much as possible. But behind this quest lies a rare complexity in terms of infrastructure. At Uber, the combination of big data and machine learning tools enables the transportation service company to analyze and understand

Figure 2. How the 'UberPool' service helps reduce traffic in the city



the behavior, locations, and preferences of its customers to more effectively manage driver availability and positioning.

The example of Uber represents a successful model of working with big data; the company does not just capture huge amounts of data from the mobile application used by its drivers and clients. His triumph is based primarily on his ability to collect relevant data to connect clients and service providers. Who needs a car, and where? It is by focusing on these two data that Uber has managed to make taxis obsolete. Uber's case needed to know exactly where the potential clients were to automate the decision-making process when sending the drivers.

GOOGLE, KNOW ALL OUR ACTIONS ON THE WEB

Big data is not new, but the rapid rate of adoption in recent times may make it appear so (Theobald, 2017). Since its inception, big data has been used by businesses to stay informed about the latest trends in sales or consumer habits. Any activity carried out on the Internet leaves an exploitable trace and in turn, produces new information. From a simple download to an online call to a click on a link, all this allows companies to determine a trend.

You thought that Google was just a search engine always working in the same way: keyword (1) + keyword (2) = search? Well think again, Google can be used in a much-targeted way by knowing the shortcuts to integrate

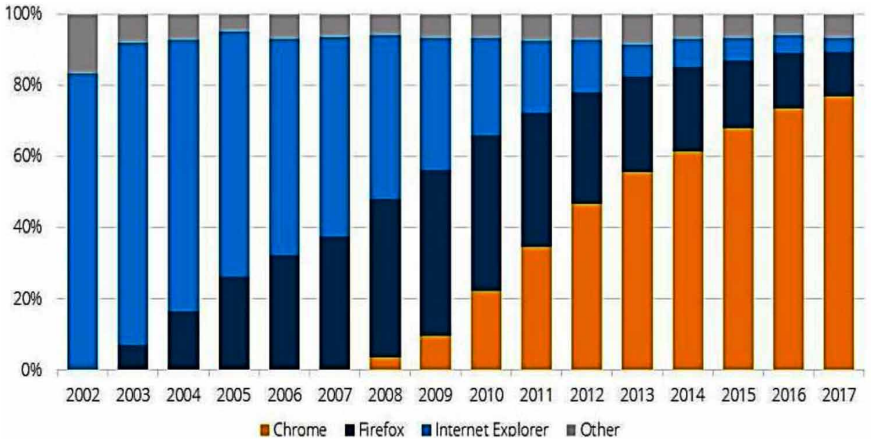
into the research. Google legitimately uses the amount of data available on the web to make the search engine the ideal companion of the user. A simple search can retrieve nearby movie schedules, transportation program, interesting restaurants, and many other services.

As another perspective, let's look at a more mature field, that of advertising with the case of Google. His business model is data-centric, with more than 80% of his revenue being generated by monetizing his big data through the sale of targeted online advertising. How did they go to access this amounts volume of data?

- **Access to the Data Flow on Our Computers:** Google introduced the free Chrome browser in 2008. The following graph shows the evolution of the market shares of different browsers: in 10 years, Google Chrome has captured almost 80% of the market (see Figure 3).
- **Access to the Data Flow on Our Mobiles:** Google bought the Android startup in 2007 and its operating system is currently used on more than 80% of smartphones.

Chrome and Android are the interfaces that allow Google to acquire a considerable amount of data on our activities. The Cupertino's company has thus managed to obtain a market share such as it has now become unavoidable. We can even summarize Google's strategy of access to the data by the quote of Peter Thiel in his book "Zero to One" published in 2014: "competition is

Figure 3. Web browser Market Share (2002-2017)
Source: www.w3schools.com/



for losers". His premise is based on a foundational belief that: "if you want to create and capture lasting value, look to build a monopoly".

Google currently collects our personal data for the primary purpose of creating targeted advertising. However, in the future, the Mountain View giant could use this data to change the behavior of individuals, or even the entire world population to solve problems such as poverty or epidemics.

Thus, Google would become in a way the brain of humanity, while individuals would be reduced to mere pawns. This fascinating and terrifying perspective is the one described in the short film '*The Selfish Ledger'*, directed by Nick Foster of Google X for an in-house broadcast.

You may be wondering how big data coupled with artificial intelligence could be used to collect more personal data. Simply, for example, if an individual doesn't have a balance in their bathroom, they could decide to print one in 3D format, in order to collect data on his weight.

Similarly, Google could use the data collected to create huge registers passed from one generation to another, like genetic information. The artificial intelligence could also be used to control the behavior of users by sending them recommendations on their smartphones. These could include suggestions to help individuals achieve their personal goals, such as running for 30 minutes to lose weight.

Google has not only significantly influenced the way we can now analyze big data (think MapReduce, BigQuery, etc.) – but they are probably more responsible than anyone else for making it part of our everyday lives (Marr, 2015). And many companies, start-ups …had mapped out their big data road based on Google's experience in order to manipulate big data.

SPOTIFY: WHEN DATA ANALYTICS ORCHESTRATES OUR LISTENING STYLE

Since its launch in 2008 in Sweden, Spotify has widened the gap. Originally, Spotify was a large music catalog in which the user was free to browse by following suggestions via similar artists. However, to grow you often have to improve, and it is here where Spotify has well managed the game.

A brief detour through the evolution of the functionalities, proposed by the platform, is necessary in order to understand how the company has imposed itself.

First, there is the discover section dedicated to discoveries which, in the beginning, is a little messy. Gradually, it becomes more structured. As this section evolves, Spotify engineers will see, thanks to the collected data that users are browsing playlists more than sections, related artists or thematic radios. It seems that associating discovery and playlist can be an ingenious idea.

This approach is further strengthened by the data. This is probably why some of the success of Spotify is explained. The changes applied to the platform were based on the data, a listening of the behavior of the users, the way that this last one has to answer or not to the last improvements. This is a data-driven strategy. The first to test this feature were its employees.

It's also the data that has allowed them to refine some details ranging from the wording presenting the playlist to the choice of the photo. In this regard, they realized that using a personalized photo had a positive impact on user engagement (+ 17%). It is therefore neither a chance nor a question of ease if, in most cases, your profile picture is a thumbnail of your playlist of discoveries of the week. All the collected data and evolutions lead to this ultra-personalized playlist.

One of the bases of the analysis remains one of the great Internet classics used by Amazon, which can be summarized as follows: "*those who listened to this have also listened to this title/artist /playlist recently"*. Even if it's based on that, Spotify is able to extend the song you were listening to thousands of songs that you might like.

However, this will not be enough to create an algorithm as precise as that of 'weekly discovers'. The frequency of listening to songs in a playlist, songs/artists … listened to recently, the favorite musical genres (more than 1500) are all data used to refine the results. All of this data is used to create a music profile. A profile that will be, according to these specificities, associated with a cluster (grouping of similar profiles detailed in chapter 6).

It is this notion of the cluster that allows Spotify in part to identify content that is likely to please. Thanks to this, the simple notion of "who listened to this appreciate that too" is outdated. Indeed, the criteria of associations are much finer and much more precise. It is then possible to use playlists created by users, titles they have recently added and to share with other users belonging to a similar cluster.

In addition, when a song is added to a playlist from a Spotify suggestion, this allows the algorithm to validate the accuracy of its recommendations. It is also able to identify the grouping within the same playlist of several

songs. If you listen to two songs, for example, that are present in the same list, it may be that a third from the latter you also like. By multiplying this on a large scale and by linking it to a library as consistent as that offered by Spotify, it is then easy to understand the ability of the company to offer each week a list of 30 songs that will delight our ears. The human finally seems to be another heart of the artificial intelligence of the platform.

This algorithm represents a considerable competitive advantage since:

- It goes beyond the simple music catalog which represents an undeniable advantage, a real differentiation for a music streaming service.
- It guides, assists and helps the user in navigating the set of proposed songs.
- It facilitates the customer experience which allows increasing the retention of consumers.

Globally, this is a strong trend of services offered on the Web. Netflix, Amazon, Google, all use the data that Internet users disseminate on the web to adapt their offer and to offer personalized services. In the era of: "consumer is king", ever more demanding, personalization seems to be an obvious solution.

The online music service is now available in many markets and has over 100 million active users. By the end of 2016, Spotify launched the largest campaign ever made. It has been declined in 14 countries in total: France to begin, then: US, UK, Argentina, Australia, Brazil, Canada, Germany, Indonesia, Mexico, New Zealand, Philippines and Sweden for a second wave.

The idea: dig into the data to discover the strangest listening secrets of their users. The Spotify systems use a network of artificial neurons as a learning method. Much like Netflix, Spotify uses machine learning to figure out users likes and dislikes and provides them with a list of related tracks.

OTHER EXAMPLES

Globally, the success of many companies in the area of big data is still based on the same power: value creation from data (Sedkaoui, 2018a). Each company has been able to identify data deposits to turn them into value. Some other examples need to be mentioned in order to prove again the power of data.

What Does Big Data Mean to LinkedIn?

LinkedIn which was created in 2003 has achieved 3.6 billion turnovers in 2016. This is not what we call a big company. However, LinkedIn has exceeded 500 million members in more than 200 countries; with two new members who join the network every second.

In these conditions, LinkedIn faces a large volume of data to process. In fact, their information system must support 2 billion searches per year by the members, processing 75 TB of data per day and 10 billion lines per day.

By analyzing all its data LinkedIn is able for example to establish the list of the most used words by its members to describe their abilities, and these words vary from one country to another. In the United States and Canada, for example, we highlight the extent of the experience, while in Italy, France or Germany we say that we are innovative, in Brazil and Spain we are dynamic and while in Great Brittany we highlight the motivation.

LinkedIn is certainly one of the companies involved in the development of what is called today in the business world the "Data Science", which is based on know-how from computer science, mathematics, data analysis and business management. Concretely, it is a question of being able to quickly collect raw data, to explore and analyze it, to translate this data into decisional information, and thus globally to reduce the time between the discovery of relevant facts, the characterization of business opportunity and the triggering of shares.

But what does LinkedIn do with its data? It typically makes analyzes to better understand and conduct its activities, but most importantly it creates products/services based on the information it generates, either globally as with the most used words, or individually with systems of recommendations (people you may know, jobs ...). Examples of data include: identifying influencers and social trends in virality; test new products / services, new sites to maximize the impact on login activity and members' use; to understand the use of services over time depending on subscription levels, the means of connection (mobile, tablet...); provide detailed advertising revenue analysis reports; to evaluate the impact of viral marketing action; optimize the recommendations engine; to create specialized functions for services for companies (marketing, recruitment, ...).

To get these interesting results from the exploitation of its data, LinkedIn had to develop its own applications for data flow management, storage, search, network analysis, etc. and of course his own dashboards. To this end, the

company turns to the market to find the needed tools or the solutions, and we can list in a non-exhaustive way: Hadoop, Hive, Kafka, Microstrategy, Pig, Tableau software, Teradata …

The Oakland Athletics

The story of *Billy Beane* and his Oakland Athletics team has had a profound impact on the history of American baseball. In 1995, the owner of the A's franchise dies and leaves his place to Alderson, after spending huge amounts to pay his players. The directives of the new directors are clear: we must reduce the payroll at all costs. Alderson will focus his attention on sabermetrics, a statistical approach to baseball, to recruit players undervalued and therefore inexpensive for the club.

In 1998, Beane succeeds Alderson and continues his work by surrounding himself with a young Yale graduate economist, Peter Brand. Together, they will build statistical models and help the Oakland Athletics become one of the best payroll/results teams. In 2006, the team has the 24th payroll of the 30 MLB teams but finishes the season with the 5th record wins/losses. Despite a lack of convincing results (only one playoff entry in 2006), Beane has completely changed the mentality of the community. In 2009, he was named by Sports Illustrated 10th in the Top 10 general managers of the decade, all sports combined.

BlaBlaCar Develops Its Marketing Strategy With Big Data

One French unicorn, BlaBaCar developed with the big data approach in order to better predict and meet the needs of their customers and prospects. Thus, BlaBlaCar seeks to understand the behavior of users to be able to develop its services. Among other things, the analysis of the data made it possible to identify a specific need from unstructured data: some passengers prefer that the driver be a woman.

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, his work on big data consists in analyzing user behaviors, optimizing the interface...

Since its creation in 2006, BlaBlaCar has gone from being a small French start-up to becoming a major transport player in Europe. The blazing growth of BlaBlaCar is largely driven by successful marketing campaigns and the

platform's ability to build customer loyalty. Far from being a coincidence, these two assets are based on the judicious meeting of a Hadoop cluster and a big data analysis platform.

The company uses big data to optimize the performance of its customer relationship management campaigns and the ergonomics of its platform.

Ali Baba: Much More Than an E-Commerce Giant

Ali Baba is a virtual gateway for anyone who wants to sell in China: 11% of retail sales in China and three-quarters of online sales go through this e-commerce giant. Its platforms are ultra-popular: 'Taobao' dominates the private-to-consumer market by 90%, while 'Tmall', where the international brands are located, controls half of the electronic transactions between shops and individuals. The group, born in Hangzhou 19 years ago, defies the Chinese economic slowdown: when the growth of the second world economy is around 6.5%, Alibaba sees its income increase ten times faster.

Since its introduction on the New York Stock Exchange in 2014, the group has not ceased to be considered as the "Chinese Amazon". But this comparison has its limits because Alibaba is not a distributor with its own warehouses, but a marketplace linking sellers and buyers.

Exploiting digital as a lever to boost physical trade: this is the meaning of Alibaba's Retail strategy. The group of Jack Ma is essential in e-commerce in China but also is much more than that. Ali Baba is full of data: mobile payment, financial services, entertainment… which can be exploited.

For the Chinese e-commerce Group, the data change everything, because this is where the nerve of data revolution lies. With its ecosystem, Alibaba has a huge amount of data on its millions of mobile users: their buying habits, their demographics, the interests of entertainment, all hosted by AliCloud and linked only to a user account.

So far, Alibaba's success has been mainly explained by its ability, starting very early, to fill untapped consumption needs, then, with its colossal user base, to become a valuable partner for brands. It is also to better understand the tastes of its users that Alibaba invests as much in content.

Thanks to all its data, Alibaba can both offer the consumer a personalized experience and offer merchants a very fine analysis of their customers, as well as a range of services (live streaming …) to allow them to be more effective.

Mobile data must also help to revitalize good old stores. Alibaba pushes its 'online to offline' (O2O) strategy, where the distinction between physical trade and virtual commerce becomes obsolete.

Also, based on the information provided by Alibaba, the Chinese authorities have launched in 2017 an operation to dismantle a network of high-end cosmetics counterfeiters. About 4000 fake products valued at more than 2.9 million $ were seized.

After opening an online store in October 2016 dedicated to high-end skin care products, Alibaba uses big data to identify websites that offer counterfeit cosmetics, including 'La Mer', 'Jo Malone', 'CK' …

In addition, in 2017, Alibaba launched 'Alibaba Big Data Anti-Counterfeiting Alliance' with 20 brands, including 'Louis Vuitton', 'Mars' and others, to bring down fake manufacturers and sellers on its platforms. So, if data is the new oil, Jack Ma intends to be the new John Davison Rockefeller.

Criteo Makes Its Voice Heard in the Big Data Era

In another industry, Criteo is one of the French internet start-ups that is currently experiencing the most success in the world. Criteo uses big data analytics tools to respond in real-time to billions of line queries to generate personalized advertising on the internet. To deliver the best for advertisers, Criteo handles massive amounts of data with response times of just a few tenths of a second, diving in the heart of its architecture.

A pioneer of so-called 'advertising retargeting', Criteo handles massive amounts of data in response times of a few tenths of a second to provide the best service to advertisers. Everyone today knows these ultra-targeted advertising banners that follow you from one website to another, reminding you of the articles you have seen on a website and offering to return to it in order to complete your purchase.

To provide such a service to advertisers, Criteo has implemented a state-of-the-art computing architecture that is best in the field of big data. For each given user, Criteo's platform knows how to identify in a few milliseconds the most targeted advertising banner that will have the greatest impact. To achieve this, it handles up, globally, thousands of HTTP requests per second. Criteo's core business is therefore based on the exploitation of big data and the RTB technique (Real Time Bidding), based on machine learning algorithms.

Airbnb Uses Big Data to Fix the Rental Price and Better-Known Clients

To determine the rental value of an apartment, Airbnb uses an algorithm called "Aerosolve". This algorithm will take into account many variables:

the city, the month, the type of property, transportation etc. In addition to traditional variables, Aerosolve also analyzes images to determine the price.

While Aerosolve helps rental companies set their prices, Airbnb has also made available to its employees a platform to help staff make queries, so that they can make decisions. For the past few years, one-third of employees have used this platform containing structured or unstructured data, images, rental data, number of rooms, user feedback, but also external events such as a festival in a city that will rent the more expensive property.

Collaborators draw on more than 1.5 TB managed by Hadoop (HDFS) and hosted by Amazon (EC2). The goal of Airbnb is to work in real-time to detect payment anomalies and go further in the personalization of the service.

As Airbnb develops, the challenges related to the volume, complexity, and darkness of the data, often compartmentalized and without context, arise within the so-called "data-driven" companies. To overcome these problems, Airbnb has developed an internal tool for searching and discovering data: a data catalog called "Data Portal".

Snickers Measure the Hunger of Its Followers

This is a unique experience that the brand has conducted with BBDM Melbourne. They started with the following observation: when we are hungry, we are nervous. With the help of data scientists, they listed 3.000 words that we use in anger and spread them on a scale to measure several levels of nervousness.

By analyzing 14,000 tweets a day, and thanks to this scale, the brand could lower the prices of their chocolate bars for each user according to their degree of anger. He had to download an application, and thus recover his digital discount voucher. This operation gave even more notoriety to Snickers. Many conversations were launched around the brand during the campaign, dating from May 2016. The operation makes sense when we know that its slogan is: "You're not you when you're hungry".

Ikea Changes the Name of Its Furniture

Everyone knows the unpronounceable nature of Ikea furniture. The brand has decided to play on this fact, renaming some of them. For the "Retail Therapy" campaign, the brand has identified some of the most popular Google queries in some countries for couple or family issues. They then renamed some of their products with these queries.

Creative people have noticed that today, whatever our problem, we are looking for the solution on the Internet. The goal was then to give to each problem a solution, which would be an Ikea product. For example, for the query: "my husband falls asleep on the sofa", the first result that appeared was a stool of the mark. In addition to being talked about, Ikea was able to increase its sales: with the Adwords purchase, the company was able to boost its search engine optimization (SEO) and its products were the number 1 search results. On top of that, the operation was really funny!

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 machine learning to strengthen its business, but also to continue its development. Then, data, as a heart of the duo big data/ machine learning, 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 machine learning, is to enable customization, merchandising and A/B testing of new features to enhance the user experience. Machine learning 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 (Sedkaoui, 2018a).

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". Face to the web actors like Pinterest ahead of the subject, eBay also invests in visual research.

A few months ago, eBay launched 'Image Search' and 'Find it', two features in this sense. With Image Search, online shoppers enter 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. That's almost 1.1 billion references.

And Many Others

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

As a result, data has passed from being a modest and oft-discarded by-product of firms' operations to become an active resource with the potential to increase firm performance and economic growth. Literature indicates that big data can unlock plenty of new opportunities, and deliver operational and financial value (Ohlhorst, 2013; Morabito, 2015, Foster et al., 2017; McKinsey, 2016).

Other companies, which are considered as leaders in this field, have developed their strategy basing on big data analytics: Apple, Facebook, Airbnb, and Tesla …

The American company 'Harrah's' has made progress in sales of 8 to 10 percent by analyzing customer segmentation data, while Amazon stated that 30 percent of its turnover came from its engine analytical recommendations (McKinsey, 2011, 2013).

The 'ad-tech' companies such as 'RocketFuel' apply statistical and optimization techniques to determine which banner ads to display.

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 data in a matter of hour's varied (meteorological and geospatial data, satellite images ...).

The use of our data by companies can also protect us. This idea can be proved by the platform created in France by Groupama. To create this platform, the Insurance used open data. It has consolidated official government data on road accidents for the period from 2010 to 2017, which required the intervention of pampers and the police. The platform allows, for a given route, to compare the fastest route (normally given by a GPS), and the safest route. People then choose the roads on which there was the least accident between the two points.

But this action also depends on its users. Indeed, each member who chooses his trips on the platform can save it and share it on his social networks. In addition to open data, Groupama uses data from its users' journeys to make the service more efficient. It's a collaborative action. The insurer will be able to retrieve data from its clients about their behavior.

In the same field, 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 which 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 is 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 with powerful tools to optimize listings, source inventory, evaluate sales and find products on leading e-commerce platforms like 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 the 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.

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.

CONCLUSION

Born in the digital age, which they have mostly helped to shape, these companies have obviously benefited from the extreme digitization of their respective activities to the point of effectively and quickly contesting the legitimacy of more traditional actors on their land.

Needless to say, companies such as Apple, but of course Amazon, Netflix and, more recently, Spotify fundamentally upset the distribution mechanisms of services, thus profoundly affecting all economic. On another playground, Google and Facebook have radically changed the landscape of communication and advertising. Banks, according to some studies, could lose up to 10% of

their direct revenue due to the influx of new entrants, such as PayPal / eBay, Google (Wallet) in the coveted market of mobile payment estimated to nearly 700 billion dollars by 2020.

These examples share common principles: an extreme digitization of their processes, leading to an extensive use of data and analytical algorithms to experiment with new business models, well beyond their original perimeters.

REFERENCES

Babcock, C. (2006). *Data, Data, Everywhere*. Available at: https://www.informationweek.com/data-data-everywhere/d/d-id/1039328

Foster, I., Ghani, R., Jarmin, R. S., Kreuer, F., & Lane, J. (2017). *Big Data and Social Science*. Boca Raton, FL: CRC Press.

Leonard, A. (2013). *How Netflix is turning viewers into puppets "House of Cards" gives viewers exactly what Big Data says we want. This won't end well*. Available at: https://www.salon.com/2013/02/01/how_netflix_is_turning_viewers_into_puppets/

Ludloff, M. (2011). *Strata Sneak Peek: Why Nobody Does It Better Than Wal-Mart*. Available at: https://blog.patternbuilders.com/2011/01/28/strata-sneak-peek-why-nobody-does-it-better-than-wal-mart/

Marr, B. (2015). *Data Strategy: Beyond the big data buzz: how data is disrupting business in every industry in the world*. Academic Press.

McKinsey Global Institute. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

McKinsey Global Institute. (2013). *Big Data, Analytics and the Future of Marketing and Sales*. New York: McKinsey.

McKinsey Global Institute. (2016, December). *The age of analytics: Competing in a Data-driven world*. Author.

Morabito, V. (2015). *Big data and analytics: strategic and Organizational impacts*. Springer International Publishing. doi:10.1007/978-3-319-10665-6

Ohlhorst, F. (2013). *Big Data Analytics: Turning Big Data into Big Money*. John Wiley & Sons, Inc.

Sedkaoui, S. (2018a). *Data analytics and big data*. London: ISTE. doi:10.1002/9781119528043

Sedkaoui, S., & Monino, J. L. (2016). *Big data, Open Data and Data Development*. New York: ISTE-Wiley.

Theobald, O. (2017). *Data Analytics for Absolute Beginners*. Kindle edition.

Thiel, P. (2014). *Zero to One: Notes on Startups, or How to Build the Future*. New York: Crown Publishing Group.

KEY TERMS AND DEFINITIONS

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): 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: 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.

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 increasing 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."

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.

Data Analysis: This is a class of statistical methods that make 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 Lake: Is a collection of storage instances of various data assets added to the originating data sources. These assets are stored in a near-exact, or even exact, a copy of the source format. The purpose of a data lake is to present an unrefined view of data to only the most highly skilled analysts, to help them explore their data refinement and analysis techniques independent of any of the system-of-record compromises that may exist in a traditional analytic data store (such as a data mart or data warehouse).

Data Science: It 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.

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

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

Proof of Concept (PoC): Is a realization of a certain method or idea in order to demonstrate its feasibility or a demonstration in principle with the aim of verifying that some concept or theory has practical potential. PoC represents a stage during the development of a product when it is established that the product will function as intended.

Section 2
The Hello World of Big Data Analytics

After understanding big data, companies move on to their analysis. Analytics help companies to (1) discover what has changed, (2) anticipate what will change, and (3) will become more proactive in making business operations decisions.

As discussed in the previous section, data is increasingly unstructured and comes from sources and takes different formats. Depending on the case, data are more or less complex to handle and frequently require real-time analysis. In this context, this section opens the second part of the book related to data analytics, which mainly focuses on big data architectures and methods. It provides a good source of advanced techniques, especially in the context of machine learning algorithms and their diverse applications. Indeed, this second section focuses on the aspects of data analytics, so the questions of data manipulation with analytics techniques, algorithms, and technology are covered. It will go from traditional analysis to the main principles of the different advanced big data analysis (Chapter 4). Then, it offers some practical advice to better parameterize the process of value creation (Chapter 5) by briefly introducing some alternative tools (data visualization and data governance) as well as the advanced technology to boost this process. Finally, this section is meant to be a panorama of different methods and algorithms that can be used when working with big data (Chapter 6).

Chapter 4
First of All, Understand Data Analytics Context and Changes

ABSTRACT

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 can be a new vector for performance. It must be recognized that without an adequate analysis, our data are just an unusable raw material. In this context, the traditional data processing tools cannot support such an explosion of volume. They cannot respond to new needs in a timely manner and at a reasonable cost. Big data is a broad term generally referring to very large data collections that impose complications on analytics tools for harnessing and managing such. This chapter details what big data analysis is. It presents the development of its applications. It is interested in the important changes that have touched the analytics context.

INTRODUCTION

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 Scandal in Bohemia)

DOI: 10.4018/978-1-5225-7609-9.ch004

Recognizing the big data universe, the opportunities and challenges and the different types of data, its significance and where to look for it, understanding the big data importance and realizing why loads of attention have been paid to the 'data revolution' were the mission of the previous section.

But, faced with the volume and the diversification, of data available today, it is essential to develop techniques to make the 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 the infrastructure of data but also about business intelligence and analytics.

Applying big data analytics is not about only knowing the R or Python language, or masters the big data technology... It is mainly about knowing why and how applies the different technical tools. 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 the big data requires an investment in computing architecture to store, manage, analyze, and visualize an enormous amount of data.

Today, companies are no longer wondering what a big data strategy can bring them. It is about knowing how to orchestrate and integrate the technological bricks. In short, the speech is also upscale from a technical point of view.

But, that's not all, because the emergence of 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 this chapter to enrich more your understanding about big data context its development, its changes: from descriptive to predictive to perspective and advanced analytics and the promises that it holds.

Before breaking down the process of data analytics in chapter five, and in order to understand big data analytics, it's necessary to look at what it is and under which circumstances it fall. That's what will be illustrated in this chapter.

BIG DATA ANALYTICS: FROM DESCRIPTIVE TO PREDICTIVE AND PRESCRIPTIVE ANALYSIS

In order to understand big data analytics, it's necessary to look at what it is and under which literature it fall. Many terms in business literature are

often related to one another: 'analytics', 'business analytics', and 'business intelligence' (BI). Davenport and Harris (2007) 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 be defined also as (Schniederjans et al, 2014):

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.

Business analytics begins with a data set 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 (Sedkaoui, 2018a).

Stubbs (2011) believes that Business Analytics goes beyond plain analytics, requiring a clear relevance 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. This data is then analyzed and mined using statistical methods and well-established techniques classed as data mining and data warehousing (Chen et al, 2012). Such type of analytics allows businesses to perform two main types (Delen & Demirkan, 2013):

Descriptive Analytics or Focuses on Reporting on What Happened in the Past

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 data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems, is accommodated in this category (Sumathi & Sivanandam, 2006).

Descriptive models can give a clear explanation why event behaved, how why certain occurred, but all this already is past perfect. So, companies can have a clear vision, based on the past, on the future, on what is more important and how they can function. This appeals predictive models which are seen as a subset of data science (Waller & Fawcett, 2013; Hazen et al, 2014).

Predictive Analytics: The Use of Past Data to Try and Predict Future Events

Liu and Yang (2017) 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 …), statistical methods, and various algorithms of machine learning 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?

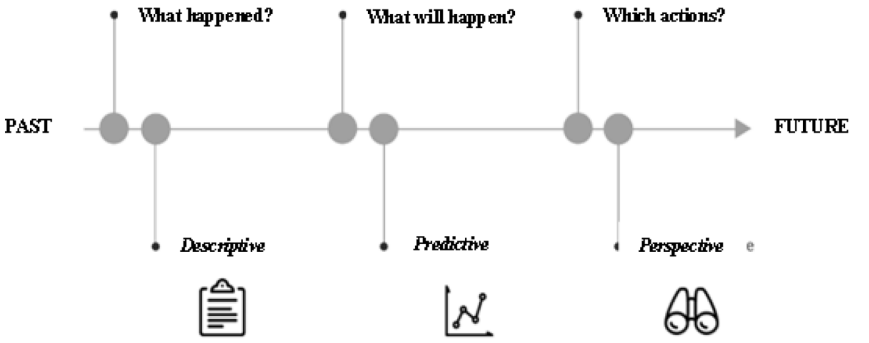
Predictive analytics use data, statistical algorithms, and machine learning to predict the likelihood of business trends and financial performance, based on their past behaviors. They bring together several technologies and disciplines such as statistical analysis, data mining, predictive modeling and machine learning technology to predict the future of businesses.

With the increasing number of data, computing power and the development of artificial intelligence 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 (Nyce, 2007; Shmueli, 2011). It has a wide range of applications in different fields, such as finance, education, healthcare, and law (Sas, 2017; Sedkaoui, 2018a).

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.

Figure 1. From descriptive to predictive and perspective analytics



And Perspective Analytics

Delen and Demirkan (2013) noticed 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 in certain time to achieve the desired outcome. So, there are many types of analytics, (see Table 1) and there is a need to organize these types to understand their uses.

When executed right, this application of mathematical and computational algorithms enables decision-makers to not only look into the future of their own processes and opportunities, but 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 machine learning algorithms and a clear feedback mechanism.

So, the application of analytics can be divided, into three main categories, namely descriptive, predictive, described previously, and prescriptive analytics.

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

Some companies may only use descriptive analytics to provide information on the decisions they face. Others may use a combination of analytic types to glean insightful information needed to plan and make decisions (Sedkaoui, 2018a).

Then, the process of analytics can involve any one of these types, the major components of business analytics include all three used in combination to

Table 1. Analytics type

Type	Description	Example
Descriptive	The application of simple statistical techniques that describes what is contained in a dataset or database.	An age bar chart is used to depict retail shoppers for a department store that wants to target advertising to customers by age.
Predictive	An application of advanced statistical, information software, or operations research methods to identify predictive variables and build predictive models to identify trends and relationships not readily observed in a descriptive analysis.	Multiple regression is used to show the relationship between age, weight, and exercise on diet food sales. Knowing that relationships exist helps explain why one set of independent variables influences dependent variables such as business performance
Perspective	An application of decision science, management science, and operations research methodologies (applied mathematical techniques) to make the best use of allocable resources	A department store has a limited advertising budget to target customers. Linear programming models can be used to optimally allocate the budget to various advertising media

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 shown in Table 1.

A process beginning with business-related data collection and consisting of the sequential application of descriptive, predictive, and prescriptive major analytic components, the outcome of which supports and demonstrates business decision-making and organizational performance, (Schniederjans et al, 2014)

Davenport and Dyché (2013) divide the ages of BI into three eras; from analytics 1.0, which started in the 1950s, to Analytics 2.0, which is when big data was firstly introduced in 2005, to Analytics 3.0, which is where are at today, and can be described as Analytics 1.0 and 2.0 combined. The big differences between 1.0 and 3.0 are: (i) the use of external data, (ii) the use of unstructured data and (iii) the use of perspective analysis.

BEFORE AND AFTER BIG DATA ANALYTICS

Before the era of IT tools, company data has been mainly handwritten paper records, not easily accessible. Now, with advanced technology, larger data amounts to be collected, stored and reused are allowed. A few years later, a new IT-term is born already the Internet of Things (IoT), in which everything

is connected. Therefore, this expanding the amount of data and consequently, increasing the importance of "Data Analytics" (Sedkaoui, 2018a).

Data analysis came in in the 20th century when the information age really began. Zhang has 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 was sparked the true revolution in data analysis.

Davenport (2014) states that company managers have been familiar with using traditional data analysis to support decisions since 1970. Vasarhelyi et al. (2015) state that traditional accounting data in companies have been ERP data, which was acquired manually in transactions.

However, the importance of data analysis started in late 1960 when researchers begin to speak about databases as repositories of data. The first databases were established from file control services in the early '60s. They were extensive and expensive program systems which were run on largely sized computers. The first significant usage fields of them were the systems in which huge pieces of data were stored, including numerous queries and modify. For example, Companies' card indexes, banks' systems, flight booking systems.

Since 1970, the publication of Codd's article, in which he suggested that the database management systems should present the data in tables for the user, database management systems have appreciably changed. E.F. Codd (Codd, 1970) 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 the database was developed. SEQUEL (Structured English Query Language) (Chamberlin and Boyce, 1974), later called SQL for copyright issues, was the forerunner of all the query languages becoming the standard for the relational database.

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. Edgar F. Codd was the first to work on data organization by designing database management systems (DBMSs), 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 academies systems, then with independent developers for free initiatives or personal, such as software, websites, etc. Even for small needs,

embedded or local systems like SQLite (http://www.sqlite.org/*)* are widely used (Sedkaoui, 2018a).

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

A management database, for example, used in ERP (Enterprise Resource Planning), 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, thanks to a good data modeling, each of these operations is optimized

But what about the statistical needs? How respond to requests for dashboards, historical analysis or even prediction?

In an ERP, this can be a complete analysis of sales trends by product category, by branch, by the department, by month, by customer types, calculating developments to determine which product categories are changing, in which region and for which customer, etc. (Sedkaoui, 2018a).

In this kind of query, or what we call "OLAP" (Online Analytical Processing), which has to cover a large part of the data to calculate aggregates, the relational model and the query optimizers of the databases can't 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 "Business Intelligence" (BI).

Data processing and analysis, in the present day, are brought together under this notion of BI, due especially to computers' increased processing capabilities. According to Chen et al (2012), the term BI became popular in the 1990s, with the term "business analytics" added in the late of 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 and more). This model, which has also been formalized by Codd, prefigures the big data phenomenon.

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 (Sedkaoui, 2018a). 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 for extracting 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 (Mitra et al, 2002). Powerful analytics tools can then be used to process the information gathered in large sets of structured and unstructured data.

In recent years, with the advent of Web 2.0 and the Semantic Web era, data analysis have become very important, replacing the traditional storing systems in many applications.

They represent now the new technology for knowledge representation, data storage, and information sharing.

Artificial intelligence and machine learning 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 characteristics (Sedkaoui, 2018a):

- Understanding unstructured data;
- The ability to extract information and derive insights;
- The ability to refine expertise with each interaction;
- The ability to see, hear and speak in order to interact with humans in a natural way.

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

Table 2. The difference between traditional and big data analytics

Criteria	Big Data Analytics	Traditional Analytics
Data types	Structured, Semi-structured and unstructured formats	Rows and columns (structured)
Data volume	100 terabytes to several petabytes or even zettabytes	Dozens of terabytes or less
Data availability	constant flow	Static
Method of analysis	Advanced Analytics (Data Mining, algorithms, machine learning …)	Based on assumptions (traditional)
Objective	Data-oriented products	Support for internal decisions and services
Technology	A stack of tools that enables to build a framework that allows extracting useful features from a large dataset	Relational databases, data warehouses, dashboards
Processing	A lot of parallel processing often strains supporting systems Can be truly ground-breaking for the organization's as previously completely unknown gaps of information can be revealed randomly rather than just providing information about what is known	Is appropriate for the analysis of data containing information that will answer to information gaps that are known
Skills	Advanced: analytical, mathematical and statistical knowledge required to develop new models (Data Scientist)	Basic knowledge of reporting and analysis tools

FROM TRADITIONAL ANALYTICS TO BIG DATA ANALYTICS

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 (Sedkaoui, 2018a).

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 (EDA) or through statistical decision models (Piegorsch, 2015). For example, achieving this goal is quite challenging to gather all the data for perfect information.

Big data are often too large for data analysts to view and process on-hand. The need for more advanced visualization techniques, capabilities to find patterns in the complexity of data and modeling capabilities have increased

along the introduction of big data (Schlegel, 2014; IBM, 2012). So, the key question is what is the difference between the two?

In conclusion, there are real differences between traditional analytics and big data analytics, and the Table 2 can help you to determine what is changing with big data analytics. In traditional statistics, there are limited amounts of data, and it must get as much information as possible out of it. 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.

Traditional analytics (descriptive) provides a general summary of data while big data analytics deliver deeper data knowledge. Traditional analytics mines past data to report, visualize and understand what has already happened? While the analytics in the big data age leverages past data to understand why something happened? Or to predict what will happen in the future across various scenarios?

A big data analytics involves the analysis of large, complex and often semi-structured and unstructured data to discover useful information and extract value. Big data analytics represented as prescriptive analytics, described at the beginning of this chapter, determines which decision and/or action will produce the most effective result against a specific set of objectives and constraints.

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 (2013a; 2013b) 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. (2011) 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 a 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.

In essence, it involves exploring a vast collection of atypical data with the aim of detecting trends and information previously invisible and achieving business objectives. This process allows businesses to go beyond what is possible with standard database systems, organized in 'rows and columns', such as viewing unstructured posts on social media or browsing through endless system logs to reveal hidden customer preferences regarding brands.

New analytics approach in big data age combines predictive and prescriptive analytics to predict what will happen and how to make it happen? Analytics uses and applications improve the efficiency of the decision-making process and generate value.

The difficulty of transforming big data into value or knowledge is related to its complexity, which is growing with the increase of its quantity its velocity and diversification of its types and sources (Sedkaoui & Gottinger, 2017).

Leveraging leading tools and techniques help to manage and extract relevant data from big data. Big data analytics can range from historical reporting through to real-time decision support for organizations based on future predictions.

NEW STATISTICAL AND COMPUTATIONAL PARADIGM

Statistics is the traditional field that deals with the quantification, collection, analysis, interpretation, and drawing conclusions from data. Development of new analytical methods is an interdisciplinary field that draws on computer sciences, artificial intelligence, machine learning, and visualization models and so on.

There are several methods that are recently developed and 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 being developed to find and design tools that explicitly reveal tradeoffs relating to complexity, risk, and time.

Concerning statistical methods literature summarizes the change in two points (Sedkaoui & Gottinger, 2017):

1. **The New Approaches Are at the Crossroads of IT Tools and Statistics:** It's concerning machine learning, where algorithms generate alone, more or less 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 pose 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 on streaming data, statistical analysis with multiple kinds of data, data quality and complexity, protecting, privacy and confidentiality, high dimensional data …

As the volume of data grows, so do the requirements for more advanced data warehouses and dispersed cloud-based databases (Kimball & Ross, 2011).

In the case of data analytics, we analyzed the requirements regarding:

- **Data:** Types, structure, format, and sources (IoT data); and
- **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 & Brynjolfsson (2011) 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's 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 (Sedkaoui & Gottinger, 2017). More frequently than not also good theoretic insights and models of the subject discipline would be helpful to identify the 'payoff relevance' of data for predictive purposes (Harford, 2014).

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, are considered as an umbrella concept for the analyzing of data with the explicit aim of generating value, in the form of efficient information, that aid the decision-makers in their process. This idea can be formalized by Van Barneveld's et al, (2012) definition.

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

While:

- 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 (Cooper, 2012).
- *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 (2012) emphasis on a problem definition with what Labrinidis and Jagadish (2012) conceptualize as data management which includes:
    - **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, increase considerably. The increasing computerization of all types of processing implies an exponential multiplication of this volume of data which counts now in PB, EB, and ZB. Chen et al, (2012) 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 semi-structured and unstructured data require 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:** 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 (Tomar et al., 2016).
- **Interpretation:** Involves making sense of the analysis results of, and subsequently conveying that information in the most comprehensible form onwards to the relevant parties. In another word, making sense of different types of data and generate value from it, results in some form of finding.

But, it's necessary to point 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 wait for the results. These barriers can be approached either with newly developed statistical methodologies and/or computational methodologies (Wang et al, 2015). This can be clarified regarding the IT and the statistical point of view (Sedkaoui, 2018a):

- **From an IT Point of View:** Knowledge of Hadoop is highly desirable. It allows the creation of distributed applications and 'scalable' on thousands of nodes to manage the amount of data bytes (PB, ZB …). The principle is to split and parallelize (distribution) data batch task to linearly reduce the computation time (scalable) depending on the number of nodes. Hadoop becomes the mining web reference tool and 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...

Every problem, especially industrial, requires a specific approach after a search that 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 also can help scholars who are designing experiments or making causal inferences from observational data.

Also, regarding the business opportunities discussed throughout the several chapters of this book, big data technology offers a wide range of technical opportunities, this includes:

Reduce Costs

By reducing hardware and software investments: open source license, commodity hardware. For the same storage need it will cost 5 times less with a NoSQL solution than with a traditional solution thanks to the decrease in costs of licenses and tools. For data analytics, the trend is even more in favor of big data technologies compared to the traditional data warehouse and BI systems.

Improve Performance / Scalability

Scalability is often 'native' in big data solutions since it was part of the challenges to solve. Scalability allows to increased solicitation by adding machines. It is the ability of the solution to distribute the load over a set of nodes. As far as performance is concerned, big data solutions have a lot of advantages by using certain peculiarities of platforms: storage, data processing …

Reduce Time-To-Market (Variability)

With more flexibility in supporting unstructured data, existing and open source connectors …big data technologies will facilitate the addition of new data sources, the interconnection with an existing silo and the implementation of a new environment. This should make it possible to reduce the time between the emergence of an idea and its implementation.

Data Collection and Data Value

Obviously, the implementation of big data technology is primarily a desire to collect, often in a Data Lake for analysis and monetization. This is not new, but the primary characteristics of big data are the ability to process a large volume of data; the scalability of solutions and of processing diverse formats. Another advantage of big data technologies is the proposition of a complete ecosystem able to handle all data problems (storage, processing, analysis, protection ...).

CHALLENGES THAT DATA ANALYTICS SHOULD ADDRESS IN THE BIG DATA AGE

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, offer significant opportunities for value creation (Sedkaoui, 2018a).

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's the use of data that empowers decision-making. Being increasingly aware of the importance of data and information, companies are pressing to rethink the way to 'manage' (data governance), to enrich and to benefit from them. This causes two main challenges:

1. The 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. Capture, manage, combine, secure, and always take advantage of a huge amount of data is 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 govern the growing amount of data but also ensure its security.

It's to highlight that 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 span of big data.

Let's think about big data in network cyber-security, 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 becomes difficult to be handled by the traditional data processing applications which trigger the development of big data applications (Muhtaroglu et al, 2013). If big data are combined with predictive analytics, it produces a challenge for many industries. The combination results in the exploration of these four areas (Inukollu et al, 2014):

- Calculate the risks on large portfolios
- Detect, prevent, and re-audit financial fraud
- Improve delinquent collections
- Execute high-value marketing campaigns

So, there are many technical challenges that must be addressed to realize the full potential of big data. Warren et al. (2015) 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:

- **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 multi-modal and heterogeneous in terms of their formats, semantics, and velocities. Hence, data analytics expose typically 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 the 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 be therefore employed in order to take into account the noisy nature of

data. Also, data can be typically associated with different reliability, which should be considered in the scope of their integration in an analytical approach.

- **The Real-Time Nature of Big Datasets:** Big data feature high velocities and for diverse application 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 advent, Internet and several connected objects provide typically high-velocity data, which however 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 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 scope of large scale deployments. Note also that the temporal dimensions of big data can serve as a basis for dynamically selecting and filtering streams towards 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 are required prior to their operationalized deployment. Note that the specification and deployment of IoT analytics systems entail 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.

Table 3. Big data analytics pipeline

Steps	Description
Data Acquisition and Recording	It is critical to capture the context into which data has been generated, to be able to filter out irrelevant 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 for cross-checking 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 re-tracing the analysis. Rich provenance is critical in this respect.

Jagadish et al (2012) provide a comprehensive discussion of such challenges based on the notion of data analysis pipeline (see Table 3).

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

So, if big data is an inexhaustible store of information, you have to know how to sort the valid data from those that are superfluous. With billions of bytes produced daily, how is it possible? The simplest solution is the use

of algorithms or applications for analysis and statistics. Thanks to machine learning, sorting the interesting data is easier. This option makes it possible to unravel complicated data found on certain sites such as social networks.

An example of an algorithm for analyzing big data is 'Google Analytics'. This software makes it possible to follow qualitatively and quantitatively the entries on the given websites. The data collected will determine what people are looking for or the profile of visitors. This algorithm can be used also to optimize SEO.

RELATED DISCIPLINES

It is not enough to store a multitude of data within a specialized database, Data Warehouse …, it is still necessary to exploit and analyze them. This is the role of data analytics which, if used properly, will draw the lessons contained in this mass of data far too important to be satisfied with statistical tools alone.

In contrast to conventional methods of statistical analysis, data analytics is particularly suitable for processing large volumes of data. With the increase in storage capacity of computing, a maximum of information will be captured, ordered, stored and processed in the data analytics process.

Now, nothing escapes collection: behavior of consumers, like on Facebook, tweets, videos on YouTube, characteristics of products or services … and with the data analysis, these databases are exploitable. Different techniques can be used, and can be chosen according to the nature and types of the data and the objective that one wishes to undertake, namely:

- Classification and segmentation techniques;
- The principles of decision trees;
- Methods based on principles and rules of associations or analogies;
- Methods exploiting the learning capabilities of neural networks;
- And for population evolution studies, genetic algorithms;
- Naïve Bayes algorithms, time series, linear regression ...

So, data analytics process (which will be detailed in the next chapter) are supported by a range of data management and analysis disciplines, including:

Statistics

Statistics provides the theory for testing hypotheses about various insights from data. It's 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 launch results of a 6-sided die, the procedure will generally consist in making the assumption that the die is balanced (each side has a chance on six to appear at each throw), then launch 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 that we will test in turn.

Machine Learning

Machine learning is an empirical approach based on data to provide or analyze problems; it is often based on data mining algorithms known for a long time, often developed before the 2000s. Unfortunately, their performance has been good often limited by the lack of available data or IT capabilities to handle large volumes of data. Today, the increase in storage capacity and the profusion of data that they generate, coupled with more and more powerful computing resources, bring these algorithms back to the forefront.

Machine learning is a data analysis technique that teaches computers what humans are naturally capable of: learning from their experiences. Machine learning's algorithms, essentially the supervised and unsupervised algorithms, use computational methods to "learn" information directly from the data without the need to rely on a predetermined equation as a model (Sedkaoui, 2018a). Algorithms adapt and become more efficient as the number of samples available for learning increases. Algorithms of machine learning identify natural patterns in the data that generate useful information and help to make better decisions and make better predictions.

Machine learning enables the implementation of learning agents based on data mining; machine learning includes several heuristic techniques. Machine learning 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. That leads us to say that machine learning 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, and then calculate an empirical probability for each result. The higher the number of launches in the learning phase, the better the results.

You need more clarity, right?

Ok! Let's briefly illustrate what machine learning can do with a simple case, probably closer to your everyday life: "*an anti-spam filter*". At first, one can imagine that the system will analyze how you will classify your incoming emails in spam. Thanks to this learning period, the system will deduce some criteria of classification.

For example, the probability that the machine will classify an email in spam will increase if the email contains terms such as "*money*", "*free*", "win"... and the fact that the sender of the mail doesn't 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 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.

The basic idea of machine learning is that a computer can automatically learn from experience (Mitchell, 1997). Using the collected data, a machine learning 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 (Eckerson, 2007).

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

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.

Globally, data mining is associated with big data. Big data is all data that can no longer be managed manually because of its size. Data processing and analysis must, therefore, be carried out using computerized methods. Data mining is considered as a substep of the process named KDD as follow:

- The choice of the database
- Pretreatment, in order to initiate a data cleanup
- Their transformation into the proper form for their treatment
- The process of mathematical analysis (data mining)
- Interpretation of the results of the analysis

The process defines a sequence of steps (with eventual feedback) that should be followed to discover knowledge in data (see the knowledge discovery process 'KDP'). To advance successfully each step we must apply effective data collection, description, analysis and interpretation (Piegorsch, 2015). 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 towards extracting knowledge. To this end, data cleaning, learning, and visualization might be also employed.

According to the Gartner Group (2017), 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.

Text Mining

Text Mining is a branch of data mining specialized in the processing of text corpora to analyze the content and extract knowledge. Just as data mining can be loosely described as looking for patterns in data, text mining is about looking for patterns in text. Text mining emerged at an unfortunate time in history. Data mining was able to ride the back of the high technology extravaganza throughout the 1990s and became firmly established as a

widely-used practical technology - through the dot-com crash may have hit it harder than other areas (Franklin, 2002).

Text mining, in contrast, emerged just before the market crash - the first workshops were held at the International Machine Learning Conference in July 1999 and the International Joint Conference on Artificial Intelligence in August 1999 - and missed the opportunity to gain a solid foothold during the boom years.

Text mining is the analysis of data contained in natural language text. It' refers to the technique which automates the processing of large volumes of text content to extract the key trends and to statistically identify the different topics that arise.

For example, we can distinguish a complaint from a customer to a request for information, or even a spam advertising message, inspecting the turn of phrases. 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 (Sedkaoui & Monino, 2016).

Database Management Systems

The database is a kind of data collection. It stores data, which is in connection with the given task, orderly. The access to the data is also taken care of by the database. Besides, it guarantees the protection of the data and also protects the integration of the data. Those program systems which are responsible for guaranteeing access to the database are called database management systems (DBMS). Furthermore, the database management system takes care of the tasks of the inner maintenance of the database such as:

- Create a database
- Defining the content of the database
- Data storage
- Querying data
- Data protection
- Data encryption
- Access rights management
- Physical organization of the data structure

A database management system stores data in such a way that it becomes easier to retrieve, manipulate, and produce information. Including Relational Database Management Systems (RDMS), NoSQL databases, big data databases, such as the HDFS (Hadoop Distributed File System), those provide the means for data persistence and management.

Data Streams Management Systems

To tackle the aforementioned challenges in data stream processing, a specific type of systems called Data Stream Management Systems (DSMS) has evolved. DSMS have to react to incoming data and deliver results to listening users frequently (Stonebraker et al., 2005). This is also termed as the 'DBMS-Active', 'Human-Passive model' by Carney et al. (2002), while the DBMS-Passive, Human-Active model is implemented by common Database Management Systems (DBMS).

DSMS use a cycle of data monitoring, managing, and mining to accommodate complex queries in streaming applications like computational finance. The notion of time is crucial in such applications, which include activities with both immediate and historic real-time data streams. To address the challenges of real-time applications, a DSMS must support or have certain features. One is the specification of ease and completeness.

DSMSs are a perfect fit for these applications, which must query over streaming data as well as data archived for historical analysis (Chandramouli et al, 2010). Indeed, several DSMSs have already become popular, each with its own execution model and approach to stream management. DSMS Handles transient streams, including continuous queries, while being able to handle data with very high ingestion rates, including streams featuring unpredictable arrival times and characteristics.

CONCLUSION

If one is right to speak of the data as a new 'oil', to take advantage of this neo-Eldorado one must control the exploration of the potential deposits, the extraction of the available data, the exploitation of the information relevant and refining useful knowledge. The rise of the new tools at the service of companies has caused a major change in the way of apprehending and exercising business activities. Companies are now very ROI oriented, with

objective and measurable KPIs at all stages of their activities. Now that you have understood the context of the data analytics process and you have got a basics different issues that you need before approaching any process ... I Think that it's time to get into the concrete. In another word, how about seeing how data analytics process takes place?

REFERENCES

Brynjolfsson, E., & McAfee, A. (2011). *Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy*. Lexington, MA: Digital Frontier Press.

Carney, D., Centintemel, U., Cherniack, M., Convey, C., & Lee, S. (2002). Monitoring streams – a new class of data management applications. In *Proc. 28th Intl. Conference on Very Large Data Bases (VLDB)*. Morgan Kaufmann.

Chamberlin, D. D., & Boyce, R. F. (1974, April). SEQUEL: A structured English query language. *Proc. 1974 ACM SIGFIDET Workshop*, 249-264.

Chandramouli, B. (2010). *Data Stream Management Systems for Computational Finance*. Tech. report, MSRTR-2010-130, Microsoft Research. Retrieved from http://research.microsoft.com/pubs/138844/streams-finance.pdf

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *Management Information Systems Quarterly*, *36*(4).

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, *13*(6), 377–387. doi:10.1145/362384.362685

Cooper, A. (2012). What is analytics? Definition and essential characteristics. *CETIS Analytics Series*, *1*(5), 1–10.

Cukier, K., & Mayer-Schoenberger, V. (2013b). The Rise of Big Data. *Foreign Affairs*, *92*(3), 28–40.

Cukier, K., & Mayer-Schonberger, V. (2013a). *Big Data: A Revolution That Will Transform How We Live, Work and Think*. Boston: Houghton Mifflin Harcourt.

Davenport, T. H. (2014). *Big data at work: Dispelling the myths, uncovering the opportunities*. Boston: Harvard Business Review Press. doi:10.15358/9783800648153

Davenport, T. H., & Dyché, J. (2013). *Big Data in big companies*. International Institute for Analytics. Available at: http://www.sas.com/resources/asset/Big-Data-in-BigCompanies.pdf

Davenport, T. H., & Harris, J. G. (2007). *Computing analytics: the new science of winning*. Boston, MA: Harvard Business School Review Press.

Delen, D., & Demirkan, H. (2013). Data, information and analytics as services. *Decision Support Systems*, *55*(1), 359–363. doi:10.1016/j.dss.2012.05.044

Eckerson, W. (2007). Predictive analytics. Extending the Value of Your Data Warehousing Investment. *TDWI Best Practices Report*, *1*, 1–36.

Franklin, D. (2002, December). New software instantly connects key bits of data that once eluded teams of researchers. *Time*.

Gartner. (2017). Retrieved from: https://www.gartner.com/newsroom/id/3568917

Harford, T. (2014). Big Data: A Big Mistake. *Significance*, *11*(5), 14–19. doi:10.1111/j.1740-9713.2014.00778.x

Hazen, B. T., Boone, C. A., Ezell, J. D., & Jones Farmer, L. A. (2014). 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. *International Journal of Production Economics*, *154*, 72–80. doi:10.1016/j.ijpe.2014.04.018

IBM. (2012). *Global Business Services, Business Analytics and Optimization Executive Report. Analytics: The real-world use of big data*. IBM.

Inukollu, V. N., Keshamoni, D. D., Kang, T., & Inukolla, M. (2014). Factors influencing quality of mobile apps: Role o mobile app development life cycle. *International Journal of Software Engineering and Its Applications*, *5*(5), 15–34. doi:10.5121/ijsea.2014.5502

Jagadish, S. V. K., Septiningsih, E. M., Kohli, A., Thomson, M. J., Ye, C., Redoña, E., ... Singh, R. K. (2012). Genetic advances in adapting rice to a rapidly changing climate. *Journal Agronomy & Crop Science*, *198*(5), 360–373. doi:10.1111/j.1439-037X.2012.00525.x

Kimball, R., & Ross, M. (2011). *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons.

Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, *5*(12), 2032–2033. doi:10.14778/2367502.2367572

Liu, G., & Yang, H. (2017). Self-organizing network for variable clustering. *Annals of Operations Research*.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The Next Frontier for Innovation, Competition, and Productivity*. Washington, DC: McKinsey Global Institute.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Mitra, S., Pal, S. K., & Mitra, P. (2002). Data mining in soft computing framework: A survey. *IEEE Transactions on Neural Networks*, *13*(1), 3–14. doi:10.1109/72.977258 PMID:18244404

Muhtaroglu, F. C. P., Demir, S., Obali, M., & Girgin, C. (2013). Business model canvas perspective on big data applications. In: Proceedings—2013 IEEE *International Conference on Big Data*. *Big Data*, *2013*, 32–37.

Nyce, C. (2007). A. *Predictive analytics white paper. American Institute for CPCU.* Insurance Institute of America.

Piegorsch, W. W. (2015). *Statistical Data Analytics*. New York: Wiley.

SAS. (2017). *Predictive Analytics: What it is and why it matters*. SAS. Retrieved from: https://www.sas.com/en_us/insights/analytics/predictive-analytics.html

Schlegel, G. L. (2014). Utilizing Big Data and Predictive Analytics to Manage Supply Chain Risk. *Journal of Business Forecasting*, *33*(4), 11–17.

Schniederjans, M. J., Schniederjans, D. G., & Starkey, C. M. (2014). *Business Analytics Principles, Concepts, and Applications: What, Why, and How*. Pearson Inc.

Sedkaoui, S. (2017). The Internet, Data Analytics and Big Data. In Internet Economics: Models, Mechanisms and Management (pp. 144-166). Bentham Science Publishers.

Sedkaoui, S. (2018a). *Data analytics and big data*. London: ISTE. doi:10.1002/9781119528043

Sedkaoui, S., & Monino, J. L. (2016). *Big data, Open Data and Data Development*. New York: ISTE-Wiley.

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *Management Information Systems Quarterly*, *35*(3), 553–572. doi:10.2307/23042796

Stonebraker, M. (2005). One Size Fits All: An Idea Whose Time Has Come and Gone. Key note of the Conference ICDE, Tokyo, Japan.

Stubbs, E. (2011). *The Value of Business Analytics*. Hoboken, NJ: John Wiley & Sons. doi:10.1002/9781118983881

Sugiyama, M. (2015). *Introduction to Statistical Machine Learning*. Morgan Kaufmann.

Sumathi, S., & Sivanandam, S. N. (2006). *Introduction to data mining and its application*. New York: Springer. doi:10.1007/978-3-540-34351-6

Tomar, G. S., Chaudhari, N. S., Bhadoria, R. S., & Deka, G. C. (2016). *The Human Element of Big Data: Issues, Analytics, and Performance*. CRC Press. doi:10.1201/9781315368061

Van Barneveld, A., Arnold, K.E., & Campbell, J.P. (2012). Analytics in higher education: Establishing a common language. *EDUCAUSE Learning Initiative*, *1*(1).

Vasarhelyi, M. A., Kogan, A., & Tuttle, B. M. (2015). Big Data in Accounting: An Overview. *Accounting Horizons*, *29*(2), 381–396. doi:10.2308/acch-51071

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, *34*(2), 77–84. doi:10.1111/jbl.12010

Wang, X., Guo, F., Heller, K. A., & Dunson, D. B. (2015). *Parallelizing MCMC with Random Partition Trees.* arXiv preprint arXiv: 1506-03164

Warren, J., Donald, J., Moffitt, K. C., & Byrnes, P. (2015). How Big Data Will Change Accounting. *Accounting Horizons*, *29*(2), 397–407. doi:10.2308/acch-51069

Zhang, A. (2017). *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.

KEY TERMS AND DEFINITIONS

Analytics: 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 mounds 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 increasing 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): 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.

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

Data Analysis: This is a class of statistical methods that make 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 Lake: Is a collection of storage instances of various data assets added to the originating data sources. These assets are stored in a near-exact, or even exact, a copy of the source format. The purpose of a data lake is to

present an unrefined view of data to only the most highly skilled analysts, to help them explore their data refinement and analysis techniques independent of any of the system-of-record compromises that may exist in a traditional analytic data store (such as a data mart or data warehouse).

Data Mining: This practice consists of extracting information from data as the objective of drawing knowledge from large quantities of data through automatic or semi-automatic 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.

Exploratory Data Analysis (EDA): In statistics, EDA is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

Key Performance Indicator (KPI): Is a high-level measure of system output, traffic or other usages, simplified for gathering and review on a weekly, monthly or quarterly basis. Typical examples are bandwidth availability, transactions per second, and calls per user. KPIs are often combined with cost measures (e.g., cost per transaction or cost per user) to build key system operating metrics.

Knowledge: It 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 optimizes automatically through experience and with limited or no human intervention.

NoSQL: Is an approach to database design that can accommodate a wide variety of data models, including key-value, document, columnar and graph formats. NoSQL, which stands for "not only SQL," is an alternative to traditional relational databases in which data is placed in tables and data schema is carefully designed before the database is built. NoSQL databases are especially useful for working with large sets of distributed data.

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

Return on Investment (ROI): Is a performance measure, used to evaluate the efficiency of an investment or compare the efficiency of a number of different investments. ROI measures the amount of return on an investment, relative to the investment's cost. To calculate ROI, the benefit (or return) of an investment is divided by the cost of the investment. The result is expressed as a percentage or a ratio.

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.

Statistical Inference: Is the process of deducing properties of an underlying distribution by 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 data set; in other words, the observed data is 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.

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.

Chapter 5
Understanding Data Analytics Is Good but Knowing How to Use It Is Better!

ABSTRACT

Collecting the data and being able to generate value from it: this is certainly the key success factor of tomorrow's champions, one that will allow you to innovate and create new business models. Faced with the 3Vs of big data, many companies are embarking on big data projects with the main objective: generating value. The goal is to succeed, by the detailed analysis of large amounts of data, to lift the veil and discover hitherto hidden models and barely perceptible correlations, as many new business opportunities that companies must grasp. The key to the success of any big data analytics initiative is to define your goals, identify specific business questions that a suitable technical architecture will need to answer, and use the data experts to generate value from data by using specific algorithms.

INTRODUCTION

In all summaries, the problems seem simpler than they actually are.

Rollo May

DOI: 10.4018/978-1-5225-7609-9.ch005

Big data can be analyzed using software tools commonly used in advanced analytical disciplines, such as predictive analytics, data mining, and statistical analysis… Traditional BI software and data visualization tools may also play a role in the analysis process, but semi-structured and unstructured data may not be suitable for traditional data warehouses based on relational databases. In addition, these warehouses are sometimes unable to meet the processing requirements imposed by sets of big data that must be updated frequently, or even continuously.

Big data analytics is the process of examining large datasets containing heterogeneous data types to uncover hidden patterns, unknown correlations, market trends, user preferences, and other exploitable information. 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 used technologies 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 the dynamic processes.

As a result, many companies seeking to collect, process and analyze big data have turned to new technologies, including Hadoop and related tools such as YARN, MapReduce, Spark, Hive, and Pig, as well as NoSQL databases. These technologies form the basis of an open source software infrastructure that supports the processing of large and heterogeneous data sets on clustered systems.

Also, who says big data says also be able to manage data. A data approach, therefore, requires data governance that is irreproachable. If the data is not the right data, the data analytics process will follow with a lot of disappointment and failures. In addition, we must not forget that big data is a highly technical field that has developed very rapidly. The result is a lack of skills it is imperative to fill. By reading this chapter you will discover the how you can conduct a data analytics process and what you need to better guide this process.

WHEN BIG DATA, ANALYTICS AND VALUE CREATION MEET

Big data is perceived by some as "the spearhead of digital transformation". This is notably what *Klaus Schwab* explains in his book '*The Fourth Industrial Revolution'*. According to him, the digital age is centered around data: this means its access and use, in order to refine products and experiences and to converge towards a world of continual adjustment.

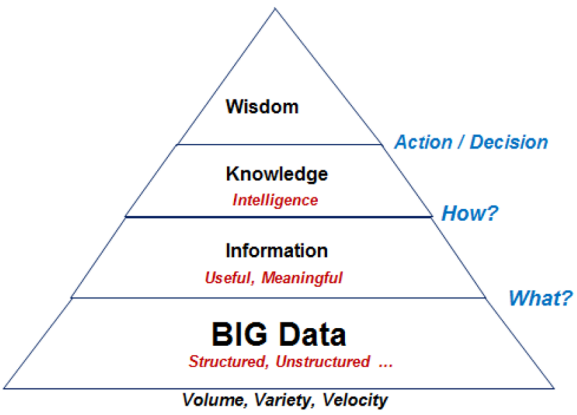
Data means the appearance of a fact which we can record, store, modify, and send on. The conception of data is not an exact idea. From the point of view of database designing, data is the meaningless series of signs from which we can earn information after processing. Ackoff (1996), defines data as symbols, information as data that are processed to be useful, and the knowledge as an application of data and information in order to have the ability to understand "what", ''why'' and ''how''.

According to Taylor (1980), 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.

Big data is everywhere, especially in the business context. The most mature companies in the exploitation of data are distinguished by the following criteria:

- The anticipation of strategic issues related to better use of internal and external data,

Figure 1. From big data to knowledge



- The diversity of data collected and collection channels,
- Creation and recruitment of data Scientists and other 'data experts',
- Adoption of new data exploitation technologies,
- Better consideration of the issues related to the protection and the privacy of personal data in the exploitation process.

Many companies have realized that knowledge is power, and to get this power they have to gather its source, which is data, and make sense of it. Data is the most valuable currency for the economics of insights, where success smiles at the companies that are faster to market their insights. Data-centric companies are making significant gains by applying continuous improvement and forward planning to improve customer satisfaction, educate key decisions, eliminate waste and reduce risk.

Turning data into information and then turning that information into knowledge (which is illustrated in Figure 1) remains a key factor for business success. With the right tools, more data can result in more knowledge. In order to understand a behavior, events, or phenomenon, it makes sense to collect all possible data that affects it. Therefore, in the era of Analytics 3.0, to collect data from external sources is fundamental. It is a way to get the more holistic view and thereby a better understanding of the behavior or phenomenon.

To ensure an efficient use of the data, it is necessary to differentiate between what we hear and what we listen to. Some data that we have are different from the ones we haven't, and from other different data formats, we unintentionally create. So, you have to know that all these data become actionable from the moment where you know what you want to do with it. The value of big data is that this phenomenon can help companies to identify useful data and transform it into knowledge by identifying models, using machine learning algorithms, technologies and new solutions (Sedkaoui, 2018a).

In this field, the Pioneers have shown the way. Data is at the heart of their business models. You have probably noticed that 'Big' companies such as Walmart, Google,

Facebook, Amazon, Apple, Amazon, IBM, Netflix, Uber, Nike and many other companies invest continuously in big data and analytics applications in order to take advantage of every data byte (Sedkaoui, 2018a).

Data can come in many forms; text, sound, video, picture, and so on. During the era of 1.0, many of these sources were not available for analysis, mainly because they were too complex. Instead, companies relied on structured and organized sources of data. Great improvements in analytics have made the more complex sources an obvious choice to include. Amounts of data

are constantly being shared and in different types, so it is not interesting to collect and analyze this data, but also to combine the different data sources in one single analysis to derive value.

Companies, regardless of their sizes and their sectors, can also consider new business models based on their data. These models can rely as much on extensions of their core business as on the creation of new products and services. Indeed, the best opportunities related to big data will probably come from the ability of companies to get out of their four walls, by considering bringing their external data information systems closer to the organization.

In this evolution, the question related to the ways that allow companies to generate value from data is imposed as an obvious necessity even as an alternative to unexpected growth. The term "value" remains rather ambiguous. For a company, it is obviously not a question of selling its files, but rather of deriving additional net income, by a profitable use of the different types, the large volume of data generated in real-time, valued by the reconciliation with other data by appealing the sophisticated analytical tools that giving them a new meaning.

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. More than anything, the key is to master the "data value chain" and to be part of this chain, an approach should be centered on their main axes. Each company can capitalize and work to organize its data value chain. Therefore, if they are best placed to know their own business, the valuation of their data assets and their exploitation is often a difficult challenge to overcome, which often requires support in the implementation of their data strategy from the first uses to its establishment.

Today, more and more companies are building business models based on data or what we call "data-driven business model". The variety of these models is important (Hartman et al., 2014; FTC 2014), reflecting many possible positions along the data value chain: collection, storage, processing, protection, and monetization. Whatever the positioning in the data chain, the question of the value of the data is rarely discussed as if it were self-evident.

Which value are we talking about? It all depends on what point of view we take: that of the customer or the company. A study conducted by Orange in Europe (Orange, 2014) illustrates the fact that customers are aware of the value of their private data without being able to really evaluate their monetary value, which is one of the challenges of the concept of self-data, and a Vendor Relationship Management (VRM) approach, in which the user manages the access of companies to their data.

Concerning the business, simply collecting data is not a source of value creation. Overall, the approach seems to be simple: (i) we need data; (ii) we need to know what we want to do with it and (iii) how to do it. But, it isn't. In addition, creating value by exploiting the data depends on many parameters.

First of all, the nature of the data (personal data, technical data, and behavioral data, primary versus secondary data) must be taken into consideration as it determines the possible treatments, in particular from a legal point of view. However, most data are valuable only if they are supplemented with metadata that describes them and enables their operation. Awareness of the economic value of metadata is recent because so far they have been perceived only for their technical role (Greenber, 2014). It is, therefore, the combination of data and metadata that will condition the value produced.

Moreover, the quality of the data is another factor that will condition the production of value, especially in the era of big data, their veracity (Teboul and Berthier, 2015).

Also, the nature of the processing (algorithmic modeling, real-time, etc.) determines the value extracted from the data. Therefore, the question of creating value from the data requires explaining these mechanisms. These are now more based on self-learning algorithms (machine learning, deep learning). Therefore, the algorithmic processing not only allows to produce a service but also to improve it continuously. The data here has a double value. In addition, the value of the data can only be conditional and not intrinsic as many analyzes suggest.

So, regardless of the value generated from the data, it remains that their ultimate value for the company will lie in the quality of their use. Reflections on use cases are in full swing to feed the necessary transformation for different sectors. Some of the data sources are easily accessible. For example, home-automation connected objects allow knowing the activity of a household: light on/off state of lamps, level of power consumption, CO2...

But these raw data without transformation are not very useful. They are a bit like wood for those who want to build a chalet. A process of transformation of the wood is essential to construct a chalet. Similarly, a big data analytics process is essential for this data to become meaningful.

The process of analyzing data is not magic. It is a process that tends to replicate human decisions, part of which is related to the combination of elements that the proliferation of new data makes legible and interpretable. The observation by the algorithms that the same circumstances produce the

same effects makes it possible to predict their occurrence (even prescribe preventive measures). Without being necessary to establish a: cause-and-effect relationship between these circumstances and these effects. This is probably what gives the illusion of magic to a completely rational process.

LEARNING TO ASK GOOD QUESTIONS

Before tackling about a data analytics process some point deserves to be relieved. I have mentioned it in my previous book, entitled "*Data analytics and big data*" published recently. To start, two essential components are needed to question whether data analytics can or cannot add value:

- Data;
- Define the business need and formalize it.

Everything in big data analytics begins with a clear problem statement. Determining what type of problem you are facing will allow you to correctly choose the technique that can be used. The success of an analytics approach cannot be possible without the clarification of what you want to achieve. This is not just valid in a big data context but in all areas. You must clearly and define what you want before undertaking anything.

So, the first of all you must imagine a path between the initial data and the value to be predicted. This means knowing what you are trying to achieve? What is needed? And why and at what level of accuracy is acceptable and actionable?

What should be done in this phase is exploring all possible paths to recover the data in order to identify all the variables that affect, directly or indirectly, the phenomenon that interests you.

In another word, some relevant questions, in this stage, are important, mainly:

- How can I make new opportunities from this data?
- Which data should I select for the analysis?
- How apply efficiently analytical techniques to generate value?
- What new insights can I expect?
- Where the greatest global potential of big data lies?
- How will these insights help me?

Have the ability to think critically allow you to understand that, big data opportunities are not in the volume of data but in the digital transformation of your business processes. For this, you need to:

Understand the Basics and Define the Task to Be Accomplished

As we have already said, data is literally the nerve 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. In another word, identify what we already know and what we don't yet?

The idea is that the more you have a good understanding of your data, the better you will be able to use them wisely during the modeling phase using several algorithms (detailed in the next chapter).

This aims to precisely determine where we should look for the data, which data to be analyzed and identify the quality of the data available but also the relation of the available data with the business problem and their meaning from a business perspective.

That means to understand what can be done with that available data before exploring it. This includes some basic knowledge about the methods that will be used and the complexities involved.

It seems obvious because data is the main raw material of an efficient data analysis process. So, if you don't 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.

Formulate some business questions to develop a method is important, such as:

- Which sources my competitors use?
- Which data should I collect to deal with their strategies?
- Why these data and how collect them?
- What can I do with these data and what results to expect?
- How much data should be processed, in another word, should I do an analysis in real-time or periodically?

So, 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 different element related to it and determine the tools you need.

It's therefore essential to have at least basic notions of statistics and mathematics to determine the right analysis technique according to the nature of each data. That means also a significant part of identifying the technologies that will be most relevant for managing the volume and the flow of data.

The specific task 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 come back in a business environment, such as product recommendations for example.

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.

Which Technology to Adopt?

Merely collecting or having access to large data sets is not sufficient to produce a result. Most of us are not sufficiently prepared for knowledge extraction process and the 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, for example, Python. To meet these information-sharing developments, we need tools across the board to help. We need infrastructure and technologies (Hadoop, Spark, MapReduce …) that accommodate ultrafast data capture and processing.

Devices, networks, central processing, and software to help us discover and harness new opportunities. The key elements to identify are: Which technology to adopt? And why should enterprises adopt this technology? Taking into account the growing advanced analytics tools.

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 Advisory*", big data technologies are adopted by 53% of companies around 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 BI.

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

Manyika et al. (2011), suggest that organizational intent should be to implement several new technologies and techniques, in accordance with the big data strategy, in order to extract value from their data.

Required Skills and Methods

It is not only a technological issue that's mater but also, how can you transform data to patters. An algorithm is a 'black box', a user can introduce data (Inputs) and he will obtain the results (outputs). How the algorithm works are not the user's business. It's like when you drive 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 …It must be considered that there is no perfect model, but models that adapt better to situations.

It must be considered that there is no perfect model, but models that adapt better to situations. Knowing some analytics methods can be a real asset (regression, classification, decision tree…). Since these different techniques can be directly implemented using the software (SAS, R, Python …) 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. Select the right method for data that you have is a very important point.

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.

What About Costs and Benefit?

First, you need to determine whether the targeted data gives you a return on the investment made by collecting and storing it. 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 like: How much data is needed for each step? What it's meant to achieve? Any presentation of data analysis results must be accompanied by a summary of the resources committed (investments…) 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 coming months. Questions that are all the more important when the company has invested in software, need for external data? Employs a consultant? Etc.

Asking interesting questions develop your inherent curiosity about data that you are working on. Knowing a little something about everything equips you to understand the context and have the ability to get out the value from data. The key is thinking broadly about how to transform data into a form which would help 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 to avoid mistakes and get the best models?
- How can you learn lessons by analyzing available data, and what are you going to do with it?
- How to best use the results of these analyzes?
- What impacts do you expect on the choices to be made?

All these questions and other questions can help you understand your data and better guide you when working with big data, this means to think about the 'meaningful' of data, so its practice.

ANALYTICS PROCESS: FROM THE SOURCE OF THE DATA TO THEIR ANALYSIS

Be careful! To give a sense to your data, it is not enough to apply an allegedly magical process to a heterogeneous set of data for relevant instruction to emerging. A big data process must go through different phases to bring out

the analytics power included in the data. When big data was born, the big analysis of the amounts data, with it different types, is required.

So, in order to help you to understand the data analytics process, I would like to remind that during the 'Taylorism period', when the most effective method for solving any complex problem consists to break it down into simpler subproblems (Sedkaoui, 2018a).

In order to take advantage and generate value from these data, it is ideal to follow an iterative process which can be structured in major steps:

Data Collection and Preparation

The most important phase of a big data project lies in the steps that precede the actual launch. We are talking about the preliminary project. This feature is not unique to big data approach. During major civil engineering works, earthworks, soil consolidation, and foundation preparation can be an important part of the project, whether in terms of costs, deadlines or technical skills to be mobilized.

When working with data, it is not the design and the modeling but the tasks of collecting, cleaning and formatting the data that will constitute the big piece, equivalent by analogy to the ground preparation work of the civil engineering works …

This preliminary phase can represent more than 3/4 of the overall project costs and deadlines. And the technical skills to put in place will not be left out. This essential phase was still rarely appreciated at its true value. His underestimation is nevertheless one of the main causes of the failures of big data projects.

Maybe you are wondering, why it is so important?

Companies are rarely aware of the heavy work that needs to be done to transform the amounts of available data into usable information to support their decision-making process. The data collected at the heart of the production systems, even when they are fully operational, cannot be used as such for decision-making purposes. The so-called production data are too often inaccurate and inconsistent or even erroneous.

On the other hand, decision analysis essentially consists in comparing data from different sources. This basic operation is only possible when the information uses the same format and in a similar logic of management.

It is simply a pity that too often we have to break our nose on this fundamental issue before grasping its significance.

Data are collected and enriched with the support of advanced technology (sensors …). 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 section one of this book), such as:

- The complexity of data (collected from different sources and different formats);
- **Noisy Data Challenge:** Big data generally include different kinds of measurement errors, outliers, and missing values;
- **Dependent Data Challenge:** In varied types of current data, such as financial time series and so on;
- 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… (Soldatos, 2017).

An absolutely essential phase, the raw data must first be valid. We analyze only usable data, i.e. 'clean' and consolidated. If you want that your data reveal the underlying message, start by cleaning your data.

A preliminary data cleaning step must be performed before any operation to correct and validate the data. In many cases, the data needs to be reworked to ensure that the amount of data is sufficient to avoid diverting the results.

So, once the data has been collected, go to the preparation step. Without lying, this is not the most enjoyable part of the process, but it doesn't make it less essential. This data preparation phase groups the activities related to the construction of dataset to be analyzed, made from the raw data. It 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 be ensured also that the data is consistent, without missing values for example. Then, all of this data must be centralized in a database.

Be sure that you don't need to know the most complex algorithms but you must have a good knowledge of the data, and prepared 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.

Analyze the Data and Build the Model

The clean data can now begin to be explored. Data analysis consists in identifying the data that have the desired characteristics and those that are divergent. It is at this stage that the power of massively parallel processing allows the exploitation of machine algorithms. These algorithms make it possible to distinguish, in the myriad of information, the weak signals representing a behavior or a situation that cannot be discovered otherwise.

This approach will be particularly useful for revealing, among populations currently difficult to identify by traditional models, for example, young motorists who have less risky behaviors. Combined with the big data, machine learning will help enrich insurance products, data analysis and new correlations throughout the lifecycle.

This step allows you 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 statistical modeling possible paths that can help to solve the problem statement.
- Introduce, if necessary, new sources of data that would help to better understand the problem.

The modeling includes the choice, the parameterization and the test 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 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.

Evaluate Your Model and Interpret the Results

Before operationalize 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 analyze.

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 back and forth between modeling phase and 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 to better understand the data.

Also, this step is crucial because it is the one that makes it possible to identify the value of data and if they have a meaning in relation to the business objective that has been assigned to the data analytics process. Indeed, a misinterpretation at this stage can invalidate a relevant piece of information and potentially lead to erroneous or less precise conclusions. It is, therefore, necessary to allocate sufficient time to the essential phases of preparation in order to qualify the data before embarking on their analysis.

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

Now Deploy the Model

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

Imagine 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 put the knowledge obtained, in a suitable form, and integrate it into the decision-making process.

The deployment can thus go, from the simple generation of a report describing the knowledge obtained to the establishment of an application, allowing the use of the obtained model, for the prediction of unknown values of an element of interest.

Finally, the cycle of the data analytics process can be schematized in the following way:

Upstream of the analytic approach there are: data and downstream refers to knowledge and then action. So I'm insisting that the importance of the data analytics process includes all the steps from recovery to deployment. A majority of the work is done even more on recovery, cleaning and data mining than on modeling itself.

DATA VISUALIZATION: DRAW ME THE DATA SO THAT I CAN SEE ITS VALUE

The collection, processing, and analysis phases, described previously, which represent a cycle of new information and which must be accompanied by big data to most effectively create value from data, need another important element which is the most transparent, intuitive, and contextual manner in which data can be used, analyzed and valued beyond simple numbers (Sedkaoui and Monino, 2016). The data itself composed of bytes and stored in different forms and cannot be invisible. To be able to see and understand these data, we need to visualize them.

Companies have an incentive to move towards analytical solutions and must complete the data value chain from the capturing of data to its presentation and dissemination; this is about another V of big data and which is called: "Data Visualization". Data Visualization provides a common language for executive directors, data scientists, and top managers, and thus, allows them to have a conversation on data.

These three groups usually have 'different business languages' and data visualization will replace these differences with a visual dialogue that everyone will understand. Data visualization is a tool for understanding the meaning of information, but it is also an instrument of communication. *William Playfair* confirms this idea by noting that:

Instead of calculations and proportions, the surest way to hit to touch the spirit is to speak to the eyes.

Also, our brains need less than 250 milliseconds to capture, understand and respond to information in visual form. Conversely, comparing several tables of raw data requires an effort of abstraction and memory that is no longer reachable from a certain volume of data. When the size of the data to be visualized increases, the classical visualization techniques are confronted with two fundamental problems, underlined by Elmqvist and Fekete (2010):

- **Perception:** The number of visual entities makes it difficult or impossible to have an overview of the data. It often happens that several data points are represented on the same pixel. This is called the occlusion problem, or over plot. The final appearance of the visualization then depends on the order in which the data is drawn. As shown by Kindlmann and Scheidegger (2014), this effect can give very different visualizations and interpretations from identical data.
- **Performance:** The amount of data to display causes a significant drop in the number of images displayed per second, which affects the interactivity of the visualization. This interactivity is essential to access the details of the data.

Interactivity is a key point of visualization. To better understand what is being observed, the company must be able to quickly change its view, in other words, direct its course of action, and access as quickly as possible clear representations of the fields to be analyzed.

Data visualization, in light of the amount of data to be displayed, is thus an important analytical tool. It is becoming increasingly clear that this is an essential aspect of the effective communication of results.

To visualize massive data, we must find techniques to overcome these two problems at once. With the rise of computing, it is possible to automate many aspects of visualization.

First, the final drawing (conversion to an image) can be done automatically. Then, the processing of the information to produce a type of visualization can be entrusted to algorithms, making the comparison of several datasets easier and faster. Finally, visualization is no longer conceived as a static image, but as an interactive system, on which the end user has feedback power.

The raw data must follow certain steps before being presented to the user, who can then act on each transformation to modify the final visualization and learn new lessons.

1. The data *preparation* stage makes it possible to choose what will be visualized. At this stage, the raw data are refined in order to keep the interesting aspects or to derive existing data values (mean, difference, map projection ...). The result of this step is a set of more or less structured data ready to be transformed into visualization.
2. The prepared data is then *filtered* to retain only a subset of interest. This filtering can be for example temporal (the result of the exploitation of a particular year), semantic (only the friends of a person of interest) or even geographical (election results on a commune).
3. Once the identified data of interest come to the step of *visual encoding* in which is decided in what form will be represented the data. This step produces a geometric representation that describes the desired visualization. It is at this stage that data is transformed into visualization.
4. Finally, the *rendering* step consists of drawing the geometry produced by the visual encoding step. This step is most often performed by a computer and may involve the use of a graphics processor. The resulting image is the final visualization that can be presented to the user.

Once you have visualized your data, you must learn from the graphs that you have created. You can ask yourself the following questions: What do I see in this picture? Is it consistent with my expectations? Are there any interesting trends? What does this mean in the context of collected data?

The data visualization can be used both as an exploratory model to find patterns and extract knowledge from processed data and as an explanatory model to clarify and illuminate relationships between data. Through the visualization of data, companies can take advantage of the real value of big data by accelerating the understanding of the available data and allow leaders to take quick and decisive action on business opportunities.

As the volume and variety of data increases, the visualization of data becomes increasingly important to stimulate a collaborative dialogue. When faced with an enormous amount of data, visual grouping can bring together points of measurement that can help decision-makers understand the relationships between data, and thus, make effective decisions.

Data visualization must be regarded as the result of a carefully considered process, which understands the capture of data quality to allow for good analysis results and the effective communication of these results throughout the process. In order to process big data more efficiently we must have visualization functions at every stage of the analytical process, namely:

- **Collection and Preparation of Data Stage:** The combination of various data sources is a key step in the data analysis process. A company must be able to visualize this process with a tool that allows for verifying that the data assembly mode accurately reflects the significance of data sources. The more a company's data source is varied, the more it needs visualization to know precisely what data it has available and to help it understand how they can help solve the initial problem.
- **Modeling Stage:** Visualization is extremely important in modeling, notably because in most cases the model must be adjusted according to the different issues.
- **Deployment Stage:** Many tools set only allow for visualization during the deployment stage. Visualization here plays a crucial role: the analysis is embedded in an application where it generates value for a variety of users and provides them with the information necessary for their work.

Quality data visualization gives managers the means to manipulate large volumes of data to bring out trends. Through dynamic comparison and cross-referencing tools, managers reveal unsuspected information that can only be revealed with massive data analysis.

With data visualization instruments, Netflix, for example, obtains data that has already been compared, sorted and put into perspective. There is only to focus on decision making. For Uber, the interface has also been redesigned to offer much more interactivity than before. Mainly for mapping elements, once the tool is configured you can sublimate all the data you have and that involves geolocation fields.

Data visualization is a growing phenomenon that we find in our applications of racing, betting or in journalism as the New York Times. The goal is always the same: to facilitate the understanding of the data so that everyone can understand them.

In this context, we can define data visualization, generally, as the formatting of raw data in an aesthetic language to facilitate its transmission, understanding, and appropriation. Data visualization can be used for different purposes, for example:

- Create and share meaningful reports with any recipient, wherever they are.
- Predict and quickly identify opportunities and anticipate future trends.

- Optimize business processes and stimulate innovation.
- Give everyone in the business the ability to visually explore and analyze all available data.

In another hand, data visualization can help you to:

- **Predict Market Trends and Grow Your Business:** Data visualization uncovers information hidden in your data and uncovers trends in your business and market that affect your bottom line. These lights allow you to gain a competitive advantage and differentiate your business activity.
- **Allow You to Know Market Dynamics Like Never Before:** By helping you to intelligently read your market, compare your position to the industry, define the most popular features of your products, and tailor their development accordingly, to link sales information with products and services, consumer preferences and much more.
- **Identify the Requirements of Each Client and Act Accordingly:** Getting to know your customer leads to better act and improved customer experience.
- **Make the Right Decisions at the Right Time, and Share Information Quickly:** Data visualization provides information that is easy to understand and share. The company's performance indicators are constantly under control. We, therefore, pass on information from the most varied internal and external sources, sharing a single source of information based on the data that will stimulate innovation.

Also, data visualization can help in dealing with data complexity. It is a technique of exploration and analysis of numerical data using graphs. This is an excellent solution to address the growing complexity of the available data, and to draw unifying lessons from it. All the protagonists interested in the subject matter are able to read and understand without ambiguity the meaning of the graphic presented if it is judiciously chosen and carefully composed.

Data visualization is considered the last kilometer of data analytics that leads us to value. It is, therefore, a post-collection stage focused on the operational, intended for the final exploitation of the data. It transforms complex data into animated graphics that are easy to understand and use. It offers an ideal alternative for sharing key information to decision-makers. It is not so necessary to be an analyst or a data scientist to understand and compile complex data.

BIG DATA TECHNOLOGY

Big data has emerged to cope with the increasing number of data. Today, the volumes of data are very substantial and it is imperative to find adequate storage and analysis solutions. The Gartner provides a very mnemonic view of the concept related to the big data, which addresses the problem of 3Vs: data volumes, data variety (variety of formats and sources) and velocity in terms of data collection, storage and analysis. And who says the appearance of a new concept says also a creation of new technologies.

So, all the power of big data is based on technology. It is the behemoths of the web like Google, Amazon, and Facebook … that have been behind the creation of these new tools.

Once a data is produced, it is necessary to be able to integrate this data into the information system of the company, then to store it in a database adapted to the type and volume of data collected. Finally, in cooperation with the business, the company will then be able to exploit the data collected via the use of distributed programming or Machine

Learning the processing of digital data masses from different channels requires specific computer tools, where most are based on the Open Source concept.

As big data continues to grow around the world, more and more companies are interested in related technologies to drive their development. The data is always more numerous, more complex, more varied, more swift and voluminous. How to exploit them? How to analyze them, how turn them into value to make better decisions? Learn about the different big data technologies that will help achieve this throughout the data analytics process:

Hadoop Ecosystem

The first and most popular solution is obviously Hadoop and its ecosystem, an open source framework, initially developed by Yahoo and now supported by the Apache Foundation for creating applications that can store and process a large amount of data in batch mode, is widely used today to handle very large volumes of data.

Hadoop implements a Massively Distributed File System (HDFS) and a Map Reduce Engine. Hadoop is supported by an entire ecosystem in order to extend its functional field, for example with HBase (NoSQL database) or Hive (data warehouse with query language to SQL). Specifically, Hadoop

consists of a part for data storage called Hadoop Distributed File System or HDFS and a part for processing information: MapReduce.

Hadoop is composed of several elements: a storage system (HDFS), a treatment planning system (YARN) and the processing framework (MapReduce). One of the best-known uses of Hadoop is the data lake. MapReduce is a massively parallel processing method and technology from the Google Corp labs with Fault Tolerance Management and the Google File System. We are talking about processing on thousands of machines distributed in clusters.

Batch Processing

They allow the data to be processed until they are exhausted at the entrance of the system. The treatments are continuous and incremental, that is to say, that the architecture will each time take into account the new data without having to treat the old ones again. To be consistent in the processing of these data, the results are visible and accessible only at the end of the treatment (once there is no data entry). It exists as MapReduce batch big data processing in its Hadoop version or Apache Spark version.

Real-Time Processing (Streaming)

This is the opposite of batch type treatments. Indeed, thanks to this method, it is not necessary to wait for the end of data processing to access the results. It is a simple solution to implement and improves processing times. They are often used as a foundation for implementing scalable solutions. Software providers today tend to offer solutions to process information quickly or even in real-time. Many projects aim to develop tools that offer the same benefits as MapReduce while being fast. Apache Foundation's Spark, Twitter Storm, and Yahoo's Software4 are just some of the many examples.

Lambda Architecture

It's a mix between batch and real-time. Using batch processing, this architecture balances latency, throughput, and fault tolerance of systems by providing accurate views of data that is simultaneously confronted with real-time data for more accurate results.

NoSQL Databases

Conventional relational databases are used to manage qualified enterprise data but are not empowered to store data on a large scale with fast processing. NoSQL databases provide a new approach to data storage that is more flexible, more scalable and less susceptible to system failures. NoSQL does not mean without SQL but refers to "Not Only SQL".

Relational databases have a very specific data organization philosophy, including the SQL query language, the transaction integrity principle (ACID), and the standardization laws. Useful for managing the company's qualified data, they are not at all suitable for very large storage and ultra-fast processing. NoSQL databases enable redundancy to better serve the needs for flexibility, fault tolerance, and scalability.

Storage 'In-Memory'

For even faster analysis, treatments directly in memory are a solution. A technology although still too expensive it is true to be generalized. Many of the leading enterprise software vendors, including SAP, Oracle, Microsoft and IBM, now offer in-memory database technology. In addition, several smaller companies like Teradata, Tableau, Volt DB and DataStax offer in-memory database solutions.

Column-Oriented Databases

Some big data software uses column-oriented NoSQL databases to maximize the flexibility of information processing. They gain performance in terms of writing and reading data, but lose in querying possibilities. However, this software finds many uses such as developing complex tariff offers or analyzing trends before launching a product or service.

HBase from Cloudera, MongoDB, and Cassandra are among the references among this software. Cassandra and HBase are basic systems of data management. These are very powerful databases for reading and writing large volumes of data. This type of database will be able to assume progressive load increments without sacrificing existing features.

Big Data Security Solutions

Because big data repositories present an attractive target to hackers and advanced persistent threats, big data security is a large and growing concern for enterprises. In the *AtScale survey*, security was the second fastest-growing area of concern related to big data. According to the IDG report, the most popular types of big data security solutions include identity and access controls, data encryption and data segregation. Dozens of vendors offer big data security solutions, and Apache Ranger, an open source project from the Hadoop ecosystem, is also attracting growing attention.

Cloud Computing

Big data requires extraordinary hardware capacity, both for storage and the CPU resources needed for processing. No need to be equipped beyond measure, the "Cloud" is there for that. The concept needs to be understood to differentiate between the private cloud and the public cloud, the internal of the external and the hybrids combining several types of solutions. Then it is also prudent to differentiate the service levels of each of the solutions: IaaS, PaaS, SaaS...

Cloud computing is not a pure big data technology, but it is the favored deployment method for big data technologies. Indeed, it requires an enormous capacity for storage and processing and the cloud is today the most capable means of supporting these volumetric and lower costs compared to a conventional on-premise solution. Also, several cloud computing tools, like Azure HDInsight from Microsoft Azure or Amazon Elastic Compute Cloud, allow you to use Hadoop to store and analyze data. On Azure HDInsight, companies are billed based on the number of nodes that are running.

Machine Learning

Machine learning is also a technology in full swing and it puts the artificial intelligence at the service of big data. These are systems that use algorithms to learn received data. In the field of big data, this technology will enable large volumes of complex data to be analyzed faster and more accurately. With improved hardware and algorithms, the results will be improved. Machine learning will also enable businesses to improve fraud detection, real-time advertising and more.

The Amazon.com recommendation engine, illustrated in big data applications in chapter 3, is one of the most representative examples of this technology. Indeed, it analyzes the profile of users, including their behavior, in order to classify them in various categories. The goal is to offer them offers that might interest them. Mahout of the Apache Foundation is today an undisputed leader in this field.

Web Analytics Software

Today, e-commerce websites or blogs have the opportunity to obtain detailed information about their visitors through specialized software. The best known is Google Analytics or Adobe Analytics which use the open source 'R' in data processing. Websites can use the reports generated by these programs to optimize their performance.

Globally, the technological creations that have facilitated the advent and growth of big data can be broadly categorized into two families: on the one hand, storage technologies, driven particularly by the deployment of cloud computing. On the other hand, the arrival of adapted processing technologies, especially the development of new databases adapted to unstructured data (Hadoop …) and the development of high-performance computing modes (MapReduce, etc.).

Therefore, there are several solutions that can come into play to optimize the processing time on voluminous databases, namely NoSQL databases (such as MongoDB, Cassandra or Redis), the server infrastructures for the distribution of the processing on the nodes and storage of data in memory.

The first solution makes it possible to implement storage systems considered to be more efficient than the traditional SQL for mass data analysis (key/value oriented, document, column or graph). The second is also called massively parallel processing. The Hadoop

A framework is an example. It combines the HDFS distributed file system, the NoSQL HBase database, and the MapReduce algorithm. Hadoop has so much in the way that almost all Data warehouse players (Oracle, Microsoft, IBM, Teradata ...) or analytics (SAS, R, Micro Strategy ...) have now announced solutions around this new ecosystem.

As for the last solution, it speeds up the processing time of requests.

Armed with these solutions and other (see Table 1), it is now possible to make sense of the big data and understand its ecosystem by bringing projects closer to the problems businesses face. Surveys of IT leaders and executives

Table 1. Big data solutions by domain

Domain	Solution
File System	HDFS: Hadoop distributed file system, Read Hat GlusterFS, Quant Cast Filesystem, Ceph Filesystem, etc.
NoSQL Database	MongoDB, ElasticSearch, Parquet, Redis, MemCache, VoldMort, Accumulo, HBase, HyperTable, Cassandra, Neo4J, etc.
NewSQL Database	Hive, HCatalog, Drill, Impala, Bayes DB, Sensei, Drizzle, etc.
Data Integration	Flume, Sqoop, Nifi, Storm, Flink, Scribe, Chukwa, etc.
Distributed Programming	MapReduce, Pig, Samza, Kudu, JAQL, Spark, PigGen, Senty, Ranger, etc.
Machine Learning	Mahout, Weka, Onyx, H2O, Sparkling, Water, MADLib, Spark, Python, R, Julia, etc.
Security	Twill, Apache Ranger, Hama, etc.
Other	Thrift, ZooKeeper, Tika, GraphBuilder, Oozie, Falco, Mesos, Hue, Ambari, etc.

also lend credence to the idea that enterprises are spending substantial sums on big data technology. "*The New Vantage Partners Big Data Executive Survey 2017*", found that 95% of companies confirmed that their firms had invested in big data technology over the past five years.

In some cases, those investments were large, with 37.2% of respondents saying their companies had spent more than100 million dollars on big data projects, and 6.5% invested more than 1 billion dollars. According to the *IDC Enterprise 2016 Data & Analytics Research*, banking, discrete manufacturing, process manufacturing, federal/central government, and professional services are among the biggest spenders. In summary, as it has already been mentioned in this chapter, it is imperative to clearly identify the problem encountered in order to be able to solve it with the help of the appropriate big data solution.

ANALYTICS SUCCESS? NOT WITHOUT DATA GOVERNANCE

The big data gets a foothold in today's world and use cases are multiplying: predictive maintenance of equipment, connected health, connected car, digitization of everyday life for all categories of employees. We are talking about big data, artificial intelligence, and machine learning the world becomes data-driven and all professionals in the sector use jargon that creates a tenacious optical illusion: everything becomes technological, information systems and

networks. Technological players are multiplying and promise professionals a total control of data flows within the company. The data will cross on a large scale to generate a value that will upset the competitive equilibrium.

But, the reality is, of course, different, because data cannot have a value without human intervention, who knows how to put it in its context of production then processing. It can only be used for a large-scale application if other people take the responsibility to monitor its quality and to implement corrective actions far beyond the computer systems. And most importantly, the data is based on an enterprise language, ontological repositories that require, to be generalized, an appropriation by the entire company of a data-oriented culture, that is to say, free from traditional silos (finance, marketing, production ...).

Far from being a simple technological fact, the transformation through the data extends and accentuates the pace of transformation of companies by the digital era. To turn the amounts of available data into value, it is necessary to invest in technology. But technology is not enough to solve a different problem related to the nature of data, their sources …

The extension of the interaction and transaction space of the Internet justifies the emergence of a new model of governance. This governance could be inspired by the dimensions studied in the IT governance (Wilkin & Chenhal, 2010; Makhlouf & Allal-Cherif, 2015): strategic alignment, the creation of value, performance measurement, resource management, risks, responsibilities or even management capacity (see chapter seven). This governance would be justified by an extension of the Internet's interaction and transaction space and by the emergence of a new 'data-driven service model'.

Collecting data, without a prior logic, without defining a specific strategy can be much less profitable than expected. Silo problems, delicate clean-up, formatting and consolidation issues, as well as the lack of competence of stakeholders to assess the importance of data, will discourage the most persistent. So, without operational "Data Governance", the ease solution will prevail. Only the most easily data will be collected, without the vision of decision-making assistance, and without added value perspective.

To ensure a relevant collection, it is essential to define the 'Why', in another word for what analysis needs, before the 'How' question i.e. what techniques, what tools. This orderly approach will allow you to ask the essential questions that are necessary to guide the project. Such as: What data do us need (collection)? What data must we archive?

What data should we secure? Etc.

So, you must be able to master this concept before embarking on or starting a big data project. The data governance lies at the heart of the digital transformation of companies.

After years of experiments that have demonstrated the interest of AI, its implementation is a mandatory step towards the industrialization of data solutions and thus the conversion of promises into earnings.

Before embarking on a big data project, companies need to implement data governance practices, with the aim of injecting quality data into their decision-making tools. The data governance is not just a technological challenge! Because, managing data means managing a wealth of knowledge and know-how that emanates from businesses activities.

So, to achieve good data governance, the following key dimensions should be considered:

First Dimension: Data Accessibility and Availability

They must be saved in the computer system and made available to users. This implies giving secure access to them and allowing them to modify the information or not while minimizing the possibility of errors. This authorization management is accompanied by a copy of the data that will allow in case of loss or malfunction of the service to maintain the quality of a service.

Second Dimension: Manage Quality and Integrity

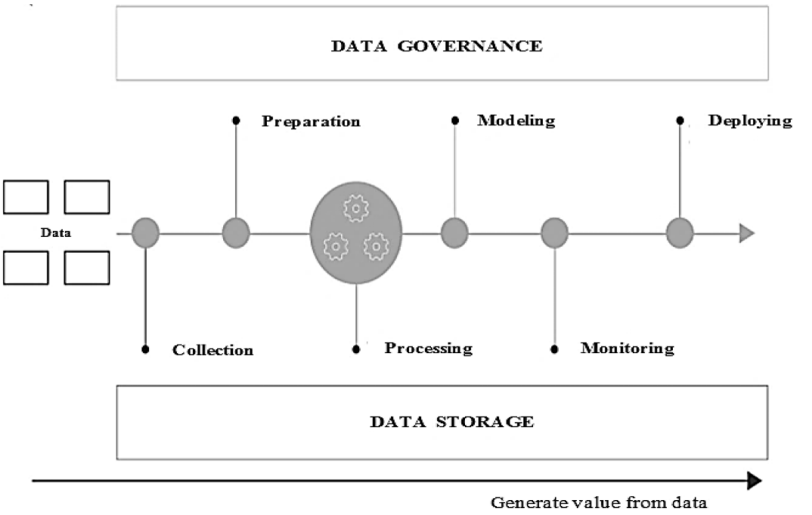
This dimension is to be correlated with availability. It involves not altering or destroying data by ensuring that it is consistent, reliable, relevant and valid. It is also a question of checking the quality of the writing on a storage medium, the transmission, etc. For this, it is essential to:

- Adopt a common definition shared by all services and management rules that ensure a sufficient level of quality for the intended use. While knowing how to prevent, detect and correct errors.
- Master the data lifecycle by identifying the source, the transformations, and the use.

Third Dimension: Privacy and Security

Data privacy and security are inseparable, at least in terms of the problem to be addressed, confidentiality issues. In all cases, the implementation of

Figure 2. Data governance



an essential data traceability program is required. Data governance involves putting in place a policy of securing the information of the company and its customers. It must make it possible to ensure the continuity of the three preceding dimensions.

Data governance is the different procedures and mechanisms set up within a company to oversee the collection of data and their use. It is as much a question of respecting the legal obligations imposed by the environment of the company, as of establishing an internal framework in order to optimize the use of the data. It is not a question of applying data governance at a product or a service level but at the dimension of the company. In any case, this is what Gartner (2016) explains in his definition of data governance.

This notion is so important that the firm offers training webinars on the subject.

So before you start a big data project, we invite you to think about many questions leading to a reflection on governance and data security:

- What are the key roles to be implemented within the companies?
- For each data perimeter, who is responsible for what; who is the Data Owner?
- What level of privacy and security, by data types?
- Which framework of rules (quality, coherence, property, documentation ...) to implement, prior to the loading of the data in a big data platform?

- Which approach should I choose to evaluate the expected level of data quality?
- Was the data useful and in which case? What joins ability with other data in the group?
- Processes to identify and manage the property (data ownership) and access to data?
- What are the processes for defining, implementing and monitoring compliance with rules and standards (quality, safety, etc.)?
- What management rules for maintaining key repositories?
- How to implement data security standards in the IT architecture?
- What tools to monitor data?

Data governance, therefore, imposes a balance between the profitability of the data collected and compliance with the rules to gain a good level of trust with the authorities and its customers. It is an important aspect of a business that aims to provide greater control over the creation, manipulation, support, storage, use, and sharing of data. A key component of any data quality or information management strategy, data governance is more than just a software solution. It requires the combination of people, rules, procedures and underlying technologies to succeed.

THE NEED FOR THE ALLY OF BIG DATA: THE DATA SCIENTIST

Big data analytics is not a modern trend, because data existing and has been used over the time and it is constantly growing. However, this progression is significantly more obvious with 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 (Sedkaoui, 2018a).

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 towards 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 the *machine* cannot solve all the problems, it makes it possible to bring some key elements of them.

Because today the technology is available, what is missing is the "Data Scientist" to exploit the data. Let's take an example often relayed on the Internet, to illustrate this point:

A few years ago, UPS conducted a comprehensive study to optimize the path of its delivery vehicles, a major challenge for a carrier that runs tens of thousands of vehicles every day. Countless parameters have been taken into account and analyzed in a conventional manner, namely structured data. The study revealed that some journeys, in all respects comparable to others, proved to be systematically more economical and faster without understanding why (the cause).

The mystery was clarified the day that data scientist realized that economic journeys coincided with those that minimized the number of turns left! Indeed, each turn to the left requires cross traffic coming from the front and therefore, there was a waiting time necessarily longer than turning to the right (priority). Thus the engines consumed fuel while idling during the wait. By calculating itineraries that favor as far as possible right-of-way turnarounds, and for 2011 alone, this helped the company to save 30 million dollars.

This example shows two essential characteristics of this new alchemist: (i) A data scientist must master data processing and (ii) also he must have a thorough knowledge of about the universe in which he operates.

A data scientist is supposed to be the expert around the data with advanced skills in three aspects. First, he must know the business context, that is to say, be able to know the problem which needs to be solved. Second, he must be able to understand how to formulate this problem in a mathematical and algorithmic way. Then get the necessary (data) and combine them to answer the relevant questions. Finally, he is supposed to make the complete implementation of the algorithm, that is to say, translate it into a computer programming language.

So, 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 taking into the account, when working with big data in a business context, is considering investment in security.

So, what it is really needed is a "data scientist"; which the *Harvard Business Review* has called *"the sexy new job of the 21th century"*; who can understand analytics and have a strong creative streak in order to ask the right questions get significant value 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 data scientist. This versatility is often represented by the "Venn diagram" (see Appendix 1) which remains a good compass to be located in this galaxy of disciplines.

The data scientist is not a statistician; he seeks to highlight characteristics to identified phenomena, which is not the case of a statistician who looks for numbers by asking specific questions.

Being a data scientist means being at the heart of data valuing and intervene 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. Then, excellent communication skills are needed.

The data scientist must seek the information where it is not. For that, the most popular way is probably to ask a lot of questions and see what sticks. But it can't be just any questions; it has to be significant questions.

CONCLUSION

The variety, volume and the velocity of data are increasing every day. The use of new storage, processing, and analysis tools has completely transformed the way information is used, especially in data analysis and preprocessing processes. Big data solutions are designed to provide real-time access to large databases.

The deployment of a big data strategy is fully in the process of digital transformation of companies, the objective being to facilitate and improve decision-making. The quest for an automated absolute prediction that eliminates all uncertainty is the ambition of the management of tomorrow. And if the algorithms are not totally autonomous, despite advances in machine learning, intelligence and human intervention are still necessary for their design and interpretation of results. For this, a particular attention must be given to the developed algorithms for big data analytics.

REFERENCES

Ackoff, R. L. (1996). On learning and the systems that facilitate it. *Center for Quality of Management Journal*, *5*(2), 27–35.

Dresden Advisory. (2017). *Big Data Analytics Market Study. 3rd annual report*. Author.

Elmqvist, N., and Fekete, J.-D. (2010). Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *Visualization and Computer Graphics, IEEE Transactions*, 439–454

Federal Trade Commission. (2014). *Data Brokers. A call for transparency and accountability*. Retrieved from https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf

Gartner. (2016). *Investment in big data is up but fewer organizations plan to invest*. Available at: https://www.gartner.com/newsroom/id/3466117

Greenberg, J. (2014). Metadata Capital: Raising Awareness, Exploring a New Concept. *Bulletin of the Association for Information Science and Technology*, *40*(4), 30–33. doi:10.1002/bult.2014.1720400412

Hartmann, P. M., Zaki, M., Feldmann, N., & Neely, A. (2014). *Big Data for Big Business? A Taxonomy of Data-driven Business Models used by Start-up Firms*. Working Paper. Cambridge Service Alliance.

IDC. (2017). *The New Vantage Partners Big Data Executive Survey 2017*. IDC.

Kindlmann, G., & Scheidegger, C. (2014). An Algebraic Process for Visualization Design. InfoVis. IEEE.

Makhlouf, M., & Allal-Cherif, O (2015). *Pertinence and Feasibility of a Unifying Holistic Approach of IT Governance*. Academic Press.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The Next Frontier for Innovation, Competition, and Productivity*. Washington, DC: McKinsey Global Institute.

Orange. (2014). *The future of digital trust. A European study on the nature of consumer trust and personal data*. Author.

Schwab, K. (2016). *The Fourth Industrial Revolution*. World Economic Forum.

Sedkaoui, S. (2018a). *Data analytics and big data*. London: ISTE. doi:10.1002/9781119528043

Sedkaoui, S., & Monino, J. L. (2016). *Big data, Open Data and Data Development*. New York: ISTE-Wiley.

Soldatos, J. (2017). *Building Blocks for IoT Analytics Internet-of-Things Analytics*. River Publishers Series in Signal, Image and Speech Processing.

Taylor, R. S. (1980). Value-added aspects of the information process. *Communicating Information: Proceedings of the 43rd ASIS Annual Meeting*, 17, 5-10.

Teboul, B., & Berthier, T. (2015). Valeur et Véracité de la donnée: Enjeux pour l'entreprise et défis pour le Data Scientist. *Actes du colloque La donnée n'est pas donnée, École Militaire*. Retrieved from https://hal.archives-ouvertes.fr/hal-01152219

Wilkin, C., & Chenhall, R. (2010). A review of IT Governance: A taxonomy to inform accounting information system. *Journal of Information Systems*, *24*(2), 107–146. doi:10.2308/jis.2010.24.2.107

KEY TERMS AND DEFINITIONS

Data Mining: This practice consists of extracting information from data as the objective of drawing knowledge from large quantities of data through automatic or semi-automatic 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: It 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.

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

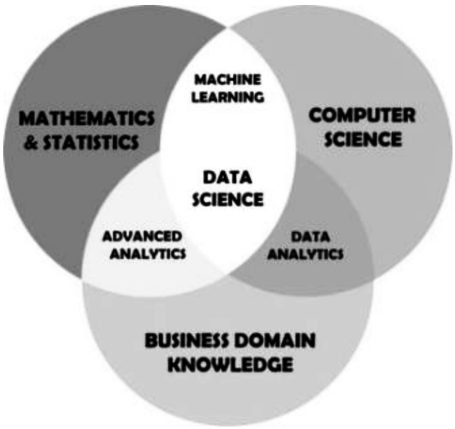
Machine Learning: A method of designing a sequence of actions to solve a problem that optimizes automatically through experience and with limited or no human intervention.

MapReduce: 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.

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

APPENDIX

Figure 3. The skill set of a data scientist



Chapter 6
Techniques and Methods That Help to Make Big Data the Simplest Recipe for Success

ABSTRACT

Data analytics has grown in a machine learning context. Whatever the reason data is used or exploited, customer segmentation or marketing targeting, it must be processed first and represented on feature vectors. Many algorithms, such as clustering, regression, classification, and others, need to be represented and clarified in order to facilitate processing and statistical analysis. If we have seen, through the 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 take a look at the different algorithms and techniques analytics that we can use in order to exploit the large amounts of data.

INTRODUCTION

Our weakness forbids our considering the entire universe and makes us cut it up into slices.

Poincaré (1913, p. 1386)

DOI: 10.4018/978-1-5225-7609-9.ch006

With the wide usage of computers and the internet, there has recently been a huge increase in publicly available data that can be analyzed. Data analyzed is 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. With big data, a new object seems indeed to have entered our lives: algorithms. Yet they have always existed: an algorithm is nothing more than a series of instructions to obtain a result. What is new is the application of the algorithm to gigantic masses of data.

More and more companies are now moving towards big data - designating the analysis of volumes of data to be processed more and more considerable and presenting a strong business challenge - in order to refine their business strategy. In addition, algorithms - complex equations programmed to perform automatically using a computer to respond to a specific problem - known only to their owners, today govern the operation of most social networks and websites.

This chapter offers a variety of methods and algorithms that can be adopted when working with big data. By opening the black box of algorithms through its categorization, this chapter helps you better understand what the algorithms do and how they work.

BIG DATA ANALYTICS COUPLED WITH MACHINE LEARNING ALGORITHMS

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 algorithms" (Sedkaoui, 2018a).

Born from pattern recognition, machine learning 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 the AI, for a game of checkers. Samul defines machine learning as the field of study aimed at giving the ability to a machine 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.

For example, if you are looking for a solution capable of estimating the houses prices. Here we can translate the definition given by Mitchell to identify: T, P, and E, so we can say:

- **Task (T):** Estimate the price of houses
- **Performance (P):** The precision of the algorithm prediction and how close it is to the real price of houses
- **Experience (E):** The description of the house and its actual price

The objective of machine learning is to learn from data, in another word, learn from real observations. As we have noticed previously, 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.

For example, by analyzing the content of the websites, 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 search phrase (Witten et al., 2016)

Another great example of machine learning applications is the IBM Watson that saved the life of a woman dying from cancer (Bort, 2016). The Watson computer system ran her genomic sequence and found it she had two strains of leukemia instead of the discovered one. This enabled another and more substantiated cure.

The goal of machine learning is training algorithms in such a way that allow them to learn from a large amount of data in order to make predictions. Depending on the type of data input, machine learning algorithms can be divided into these different families of algorithms:

Supervised Algorithm

In the case of this family of algorithms, 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 amounts of data.

Take the example of an algorithm that detects faces, a user will have to show him what a face is and what is not, so that algorithm can learn and predict if the next pictures refer to a face or not. So, it is called supervised because the learning process is done under the supervision of an output variable.

In the supervised algorithm, input data comes with a known class structure (Mohri et al., 2012; Mitchell, 1997). 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 (these two methods will be more detailed later in this chapter).

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".

The supervised algorithm can also be applied in the detection of spam from your mail but also in the forecast scores and risks associated with insurance.

To accomplish its task, this algorithm is divided into two parts:

- The first is to determine a tagged data model.
- The second consists in predicting the label of a new datum, knowing the previously learned model.

Globally, this algorithm learns from examples which need to be labeled in order to ensure the effectiveness of its learning.

Unsupervised Algorithm

In the unsupervised algorithm, input data does not have a known class structure, and the task of the algorithm is to reveal a structure in the data (Sugiyama, 2015; Mitchell, 1997).

An unsupervised algorithm will find patterns or structure in the data, so 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. The unsupervised method usually generates too many false alerts so it is often a good idea to combine both supervised and unsupervised methods as in (Krivko, 2010).

In the case of this family of algorithms, 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.

Table 1. Algorithm application examples

Question	Algorithm	Example
A or B or C …	Supervised (Classification)	To attract more customers, the best is to apply a $ 10 coupon on the purchase that exceeds $ 50 or a discount of 50%?
How much? How many?	Supervised (Regression)	How many (benefit) can the company achieve next year?
How are the data organized?	Unsupervised (Clustering)	What viewers like, what types of movies?

An example can be the discovery of different customer groups inside the customer base of the online shop. For example, if an Amazon new purchase proposal coming from you (as a new user), in such case Amazon users, are divided into groups and, according to your purchase choice, you will be associated with a group of clients who have purchased close to yours. It's just about bringing clients into groups that are not predefined.

So, the algorithm must discover by itself the pattern according to the data. The algorithm seeks to find only the similarities and distinctions within these data, and then group together those that share common characteristics.

The table 1 can help you to understand the difference between these two algorithms, by illustrating some applications examples for each one.

Reinforcement Learning Algorithm

This algorithm represents a solution perfectly suited to automated systems that must take a large number of small decisions without human instructions. For example, at home, a temperatures control system should decide if it is necessary to adjust the temperature or leave it as is. These algorithms learn on the basis of results to decide the next action.

Transfer Learning Algorithm

This algorithm 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 in some way in using the knowledge acquired to re-apply it in another environment. For example, the documents are written in two different languages.

Anomaly Detection

Anomaly detection is a machine learning algorithm for detecting abnormal patterns. This algorithm is very useful for detecting fraud in banking transactions, and intrusion detections. If you have a credit card, you have may be benefited 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 do not normally buy or an abnormally expensive purchase.

The table 2 illustrates the different questions to which these three algorithms are supposed to answer.

With this variety of algorithm, choosing the right one is a seemingly tedious process: there are many of supervised and unsupervised machine learning algorithms, in addition to reinforcement and transfer learning and each one approaches learning in a different way. Decide which algorithm to use 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.

Globally, here are some tips to help you choosing (Sedkaoui, 2018a):

- Use the supervised algorithm 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, to identify car brands appearing on video recordings of a webcam).
- Use unsupervised algorithm 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.
- Use reinforcement learning if you want to choose an action in response to each data point. Reinforcement learning algorithm is a common approach in robotics, where the game of sensor readings at a given

Table 2. The difference between the three algorithms

Algorithm	Objective	Questions
Reinforcement learning algorithm	Action to be carried out	What should I do now?
Transfer learning algorithm	Use the knowledge to re-apply it in another environment	Other languages?
Anomaly detection	This algorithm signals unexpected or unusual events or behaviors.	Is this Weird?

moment is a data point and the algorithm must choose the next action of the robot. It is also suitable for IoT applications.
- Use transfer learning algorithm if you want to use a set of tasks to influence learning and improve performance on another task.
- If you want a guide to get advice on where to look for problems then use anomaly detection.

LINEAR REGRESSION

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, the second 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: (i) classification and (ii) regression.

Regression method takes finite set relations between the dependent variable and the independent variables and creates a continuous function generalizing these relations (Watt et al., 2016). 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/is continuous, then it is a regression problem.

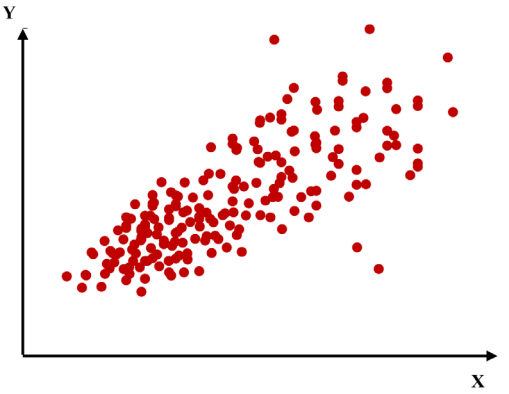
For example, you want to predict the income of clients and prospects based on data such as their socio-professional category: age, gender, occupation, address, and so on. You can collect many observations by conducting a survey on a panel of clients. Then, some of these observations will be used to generate 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 clients' income and prospects those who are not in the panel.

In the context of a regression problem, Y can take the infinity of values in the continuous set of (denoted $Y \in \mathbb{R}$). It can be temperatures, sizes, GDP, unemployment rates, incomes, or any other type of variables that do not have finite values a priori.

Regression is a simple first example of how an algorithm can learn a model. Remaining on the discussed example, the question we will try to solve now is: Given the characteristic of the bank clients' regular monthly income. In another way, how much should they normally save?

Figure 1. Savings level by income



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 1 depicts a two-dimensional graph that shows the relation between the income (X) and the dependent variable indicating their saving level (Y).

Clearly, from the first visualization of the figure, we can say that the level of savings depends linearly on the income. We can, therefore, emit 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 a form of predictions.

Then, we have our underlying model constraint which must be in this form:

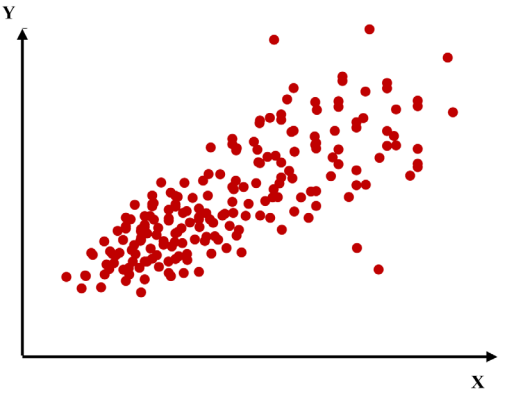
$$\hat{Y} = \beta^T x$$

With: $x = \left(1, x_1, x_2, \ldots\ldots x_n\right)$
And $\hat{Y} = \left(\hat{y}_1, \hat{y}_2, \ldots\ldots, \hat{y}_{n1}\right)$

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:

$$\hat{Y}_i = \beta_0 + \beta_1 x_i$$

Figure 2. Regression model



The goal is to find the line parameterized by:

$= (\beta_0, \beta_1)$, (this fills better the training data)

We can then graphically represent the regression equation we found to verify that it fits well with the data (observations) (Figure 2).

Now, that we have our parameter β that is to say that we have found the line that best fills our training data, we can make predictions on new data; that is to say, predict the saving by applying directly a given income as an input in the model.

But, do not forget that, in your data analytics process, you aim to find a model that approximates the reality (the phenomenon) using this model we will be able to predict.

So, before starting the phase of modeling (see the data analytics process described in the previous chapter), there are three essential elements, to take into account:

- **Description:** 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 segmentation of logic of saving, makes it possible to build a typology of clients and better target an offer.
- **Prediction:** A model can be used to anticipate future behavior. It can be used, for example, to identify the future clients of the bank.

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.

- **Decision:** 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 a discovery of a hitherto unknown characteristic concerning a phenomenon, by detection of certain profiles or the "*look-alike*". Recommendations engines prescribe scenarios of possible actions that take into account both predictions and instructions from the client.

It should be mentioned that to build your model you must minimize the "loss" of information due to the approximation already mentioned. It refers to the gap between your model, which presents an approximation of the phenomenon, and the reality. It is to say here that, the more the gap decreases, the closer we get to reality, and the better our model. But, how we can present this loss?

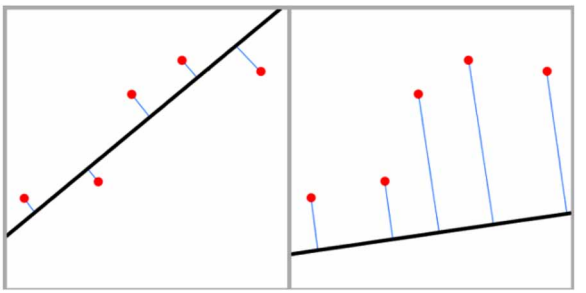
This loss can be presented by:

Minimizing the Model Error

We can represent 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 3).

On the left side of the figure, we do not lose too much information. While on the right side we are too far from reality.

Figure 3. Gap between model and reality



Maximizing the Likelihood of the Model

Indeed, the loss, in this case, is a bit hidden, but one can mathematically find that maximizing likelihood is actually equivalent to minimizing a loss function. The objective is to converge towards 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 determined a function (Sedkaoui, 2018a).

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.

CLASSIFICATION

In contrast to 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:

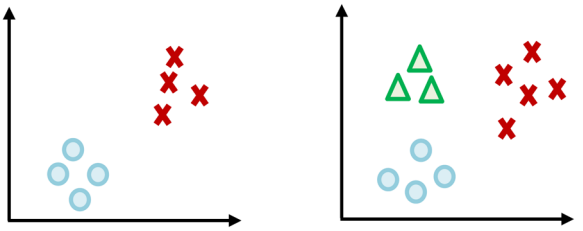
$$\left(Y = \left\{1, \ldots, k\right\}\right)$$

This is called tags assigned to the input values. This is the case of: "True/false" or "Passed/failed". This method can be used also, for example, in health risk analysis. A patient's vital statistics, health history, activity levels and demographics can be cross-referenced to score (a level of a risk) and assess the likelihood of illness.

When the set of possible values of a classification exceeds two elements, one speaks Multi-class Classification. Figure 4 illustrates both types of classifications.

Among the classification algorithms, we find K Nearest Neighbors, Logistic Regression, Support Vector Machine (SVM), Naïve Bayes, Decision Tree, Random Forest, Neural Networks…

Figure 4. Classification types (left: Binary Classification; right: Multi-class classification)



K Nearest Neighbors (kNN)

The kNN or "k Nearest Neighbors" is an algorithm that can be used for both classification and regression. The principle of this model consists in 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, Red Circle and Black Square. So, the input is bidimensional, and the target is the form to classify (Figure 5).

Now, if we have a new entry object whose class we want to predict, how could we do it? (Figure 6).

Figure 5. Simple classification example

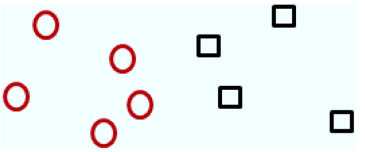


Figure 6. The blue circle is a new entry

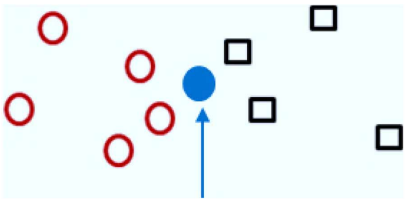
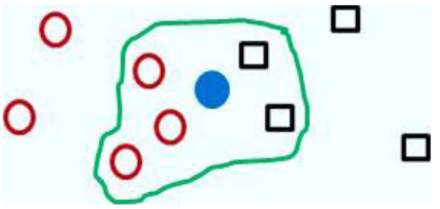


Figure 7.



Well, we'll just 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.

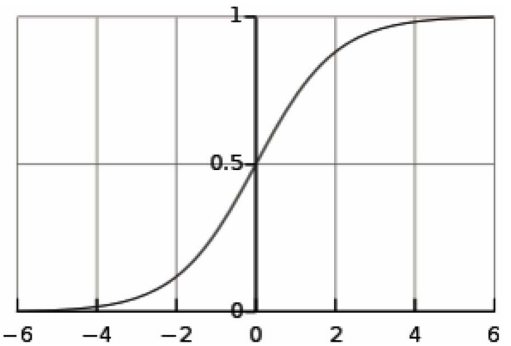
For example here, if we use the 5-NN, we can predict that the new entry belongs to the red circle class since it has 3 red circles and 2 black squares in its entourage (Figure 7).

So the principle of this algorithm is to classify a data set 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.

Logistic Regression

It's a statistical method for performing binary classifications. It takes as input qualitative and/or ordinal predictors and measures the probability of the output value using the sigmoid function (see Figure 8). We can perform the multi-class classification (for example, classify a picture in three possibilities like fruits, legumes, and roses).

Figure 8. Sigmoid function



Support Vector Machine (SVM)

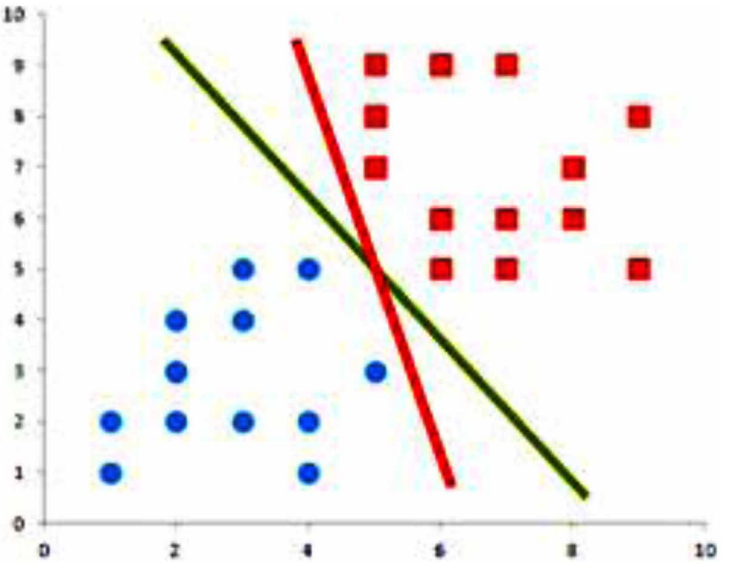
SVM is also a binary classification algorithm. If we take the Figure 9, and we consider that blue represents a class (non-spam mail for example), and red can refer to Spam. After tagging some words and concepts, the "signature" of the message can be injected into a classification algorithm to determine whether or not it's 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.

Naïve Bayes

Naïve Bayes is a fairly intuitive classifier to understand. It assumes a strong (naive) 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).

Figure 9. Example of SVM



Naïve Bayes classification is a machine learning method relying on the Bayes' Theorem:

$$P\left(A / B\right) = \frac{P\left(B / A\right) P\left(A\right)}{P\left(B\right)}$$

It can be used for both binary and multi-class classification problems. The main point relies on the idea of treating each feature independently. Naive Bayes method evaluates the probability of each feature independently, regardless of any correlations, and makes the prediction based on the 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 easiness of understanding. In addition, it performs well on the data sets with irrelevant features, since the probabilities of them contributing to the output are low. Therefore they are not taken into account when making predictions.

Moreover, this algorithm usually results in a 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 (Bishop, 2006).

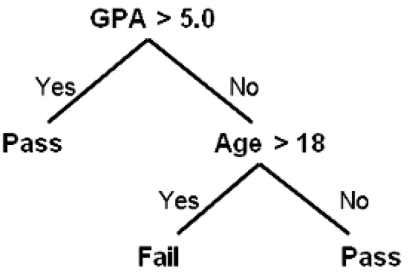
Decision Tree

Another classification method is the decision tree. Decision trees are graph structures, where each potential decision creates a new node, resulting in a tree-like graph (Quinlan, 1987). The decision tree is an algorithm based on a graph model (the trees) to define the final decision. Their purpose is to explain a value from a series of variables, so we are in the classical case of an X matrix with m observations and n variables associated with a value Y to explain.

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. The following figure illustrates an example of this operation.

A decision tree is built using a machine learning algorithm. Going from a predefined classes set, the algorithm searches iteratively for the most different variables in the classified entities. Once this is identified, and the decision

Figure 10. Decision tree example



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.

Random Forest

Random Forest is one of the most popular machine learning algorithms. It requires almost no data preparation and modeling but usually results in accurate results. Random Forests are based on the decision trees described previously (see Appendix 1). More specifically, Random Forests are the collections of decision trees, producing better prediction accuracy. That is why it is called a "forest" – it is basically a set of decision trees.

Random Forest is a combination of two algorithms: tree bagging and feature sampling. Tree bagging brings a significant improvement in performance for decision trees. Take the case of the binary classification problem; we have an X learning matrix of m learning examples, each described by n variables and a binary vector Y of m dimension. The construction of B decision trees will be as follows:

- Shoot randomly and with replacement B samples of (X, Y), which allows us to create (X_{bi} , Y_{bi}).
- Train a decision tree on each pair (X_{bi} , Y_{bi}).

For each data, we apply each of the B trees, and then we just take the majority of the B answers. In addition to the random draw, Feature sampling consists of applying a random draw on the variables to be used. The advantage of the Random Forest algorithm is the independence of the trees to be trained, which makes it possible to parallelize the treatments and improve the performance of the learning.

Neural Networks

Neural networks are inspired by the neurons of the human nervous system. They allow finding 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.

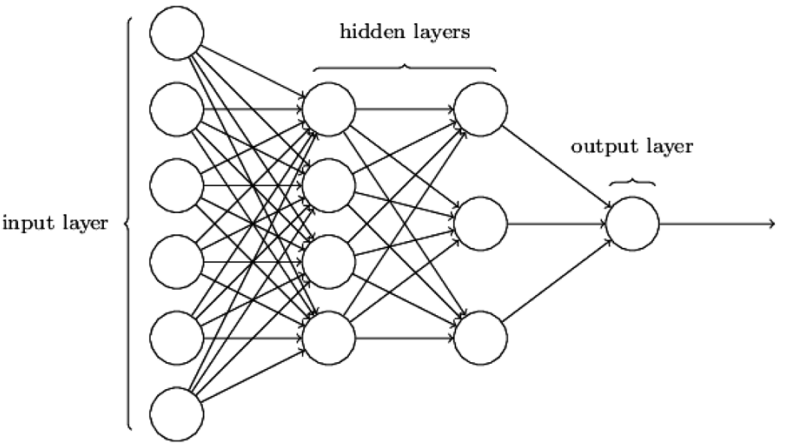
The Input Layer will then propagate the data to Hidden Layers. Finally, the third party (Output Layer) produces the classification result. Each third of the neural network is a set of interconnections of the nodes of one third with those of the other thirds.

In these networks, the learning phase aims to converge the data parameters to 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, 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 provide better predictions.

Deep learning is particularly effective in the processing of images, sound, and video. It is found in the fields of health, robotics, computer vision, etc.

Figure 11. Example of Neural networks



Each of these 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 which class a data belongs to, for example, a new email is it a spam or not?*

CLUSTERING

Machine learning is undoubtedly one of the major assets in understanding the today's and tomorrow's society challenges. Among the different components that make up this discipline, we will focus here on one of the subdomains of application that characterizes it: 'clustering'. This field covers diverse and varied subjects and makes it possible to study the associated problems according to different perspectives.

Clustering aims to determine a segmentation of the studied population without a priori on the number of classes or 'clusters' and to interpret a posteriori the clusters thus created. Here, a human doesn't 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…

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.

Find patterns in data with clustering algorithms, seems to be an amazing work, but what does it allow?

At first glance, one might think that this method has little use in real life applications. But, it is not the case, because the applications of clustering algorithm are numerous.

We have seen it in the third chapter, by wondering how Amazon (Amazon Web service) do to recommend the right products, or YouTube which can propose you several videos related to your expectations, or even Netflix that recommends good movies, all this is by applying clustering algorithm (Sedkaoui, 2018a).

The efficiency of applying a clustering algorithm can allow a significant increase in the turnover of an e-commerce site as for Amazon for example.

Also, if we provide a set of animals pictures without specifying which animals they are, then the algorithm will group, for example, the pictures of dogs together, cats together, monkey together and so on.

A cable TV that wants to determine the demographic distribution of networks viewers can do so by creating clusters from available subscriber data and what they are watching.

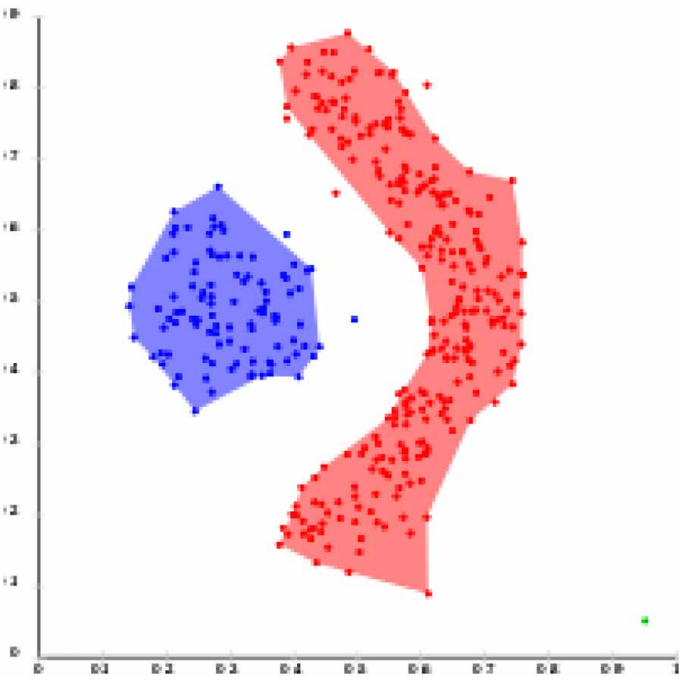
Another example can refer to the discovery of different customers groups inside the customer base of the online shop.

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, though, among other things, the grouping of differential genes according to their expression profile in a biological phenomenon over time.

Also, for the most music lovers, these algorithms can also be used, as already mentioned, for a recommendation, in order to distribute different music (Spotify for example) in clusters and to propose a song "similar" to that just listened to.

Figure 12. Clustering illustration: example



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 sub-domain of machine learning through customer segmentation, the subject of considerable importance in the marketing community.

We can think more specifically about the detection of fraud, whether in public transport, as part of a complementary health reimbursement or about the 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.

But how it works? Good question, we need more details to better clarify and understand this algorithm.

This algorithm, which is classed in the family of unsupervised machine learning algorithm, will group this data by similarity. Since we are talking about similarity, we must also talk about 'clustering'. Clustering makes 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:

$$\left\{x_1, x_1, \ldots, x_1\right\}$$

Into a set of disjoint clusters:

$$\{\left\{x_{1,1}, x_{1,2}, \ldots, x_{1,n},\right\}, \left\{x_{2,1}, x_{2,2}, \ldots, x_{1,n}\right\}, \ldots\ldots\ldots, \{x_{n,1}, x_{n,2}, \ldots, x_{n,n}\ \}\}$$

That contains 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'.

$$V = \begin{bmatrix} \\ \\ \\ v_1 \\ v_2 \\ \vdots \\ \vdots \\ v_n \end{bmatrix}$$

The number of elements of this vector is the same for all elements and it's called the 'vector dimension', denoted by n.

Given two vectors $V1$ and $V2$, we must define the distance between these two elements $d\left(V1,V2\right)$.

Typically, the similarity between two objects is defined by Euclidean distance, Manhattan distance or Hamming distance.

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:

$$d\left(p,q\right) = \sqrt{\sum_{i=1}^{n}\left(p_i - q_i\right)^2}$$

2. Manhattan Distance

The name of this distance is inspired from the famous district of New York, consisting of many skyscrapers. It is impossible to go from one point to another, it is necessary to circumvent the buildings:

$$d\left(p,q\right) = \sum_{i=1}^{n}\left|p_i - q_i\right|$$

Figure 13. How to calculate Euclidean Distance

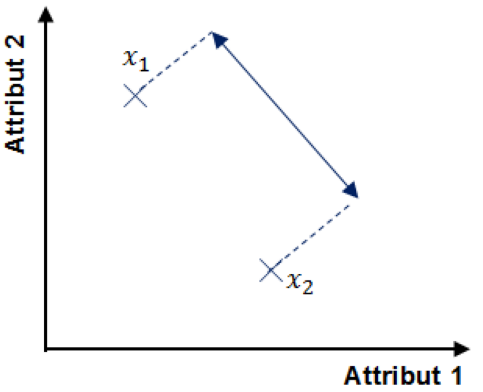
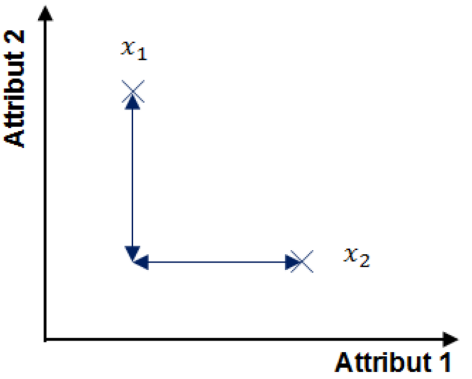


Figure 14. Illustration of Manhattan Distance



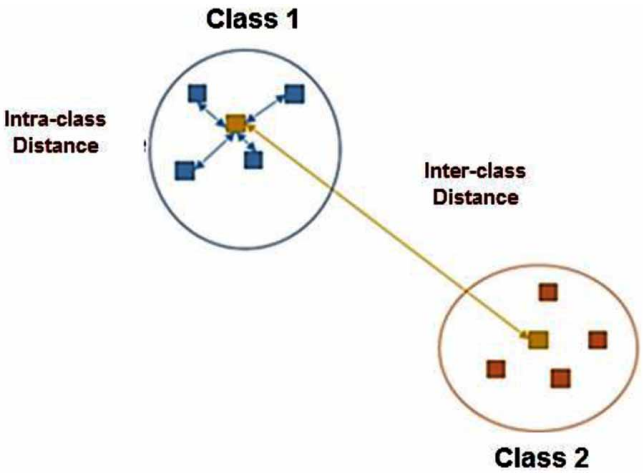
3. Hamming Distance

This distance measures the similarity between 2 words by counting the number of characters that differ.

Perhaps you realize now that there are two important notions: 'distance' and 'similarity'.

In Figure 15, you have an example of clustering. Points are grouped into two groups: Cluster 1 and Cluster 2. To calculate the similarity we chose the Euclidean distance as a metric. This metric is the simplest and also the most intuitive. In figure 15, 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' points (or 'similar') in the same group.

Figure 15. Principle clustering algorithm



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 to assign classes according to:

- Minimize the distance between the elements of the same cluster (intra-class distance).
- Maximize the distance between each cluster (inter-class distance).

However, to do this we have to try all possible combinations and choose the solution with the minimum intra-class distances, and the maximum inter-class distances. To assign classes to 27 elements optimally, it would take billion years to a processor of 3 GHz to achieve this task. This is why we use different algorithms to find the solution as close as possible. This can be the case for K-means for example, that also need to be more detailed.

K-Means Algorithm

Let's think that you look for launching an advertising campaign and that you wanted 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's what the K-Means algorithm will do!

K-means is a type of clustering algorithm 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. **First:** Since the number of classes K is imposed, this algorithm will choose K points randomly to initially constitute the representatives of each class.
2. Then, for each point this algorithm:
    a. **Calculate the distances between this point and the classes' representatives:** It begins by randomly choosing K centroids from our observations. Each point is then associated with the centroid of which it is closest.
    b. **Assign at this point the class with which its distance is minimal:** Thus forming K clusters.
    c. **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.

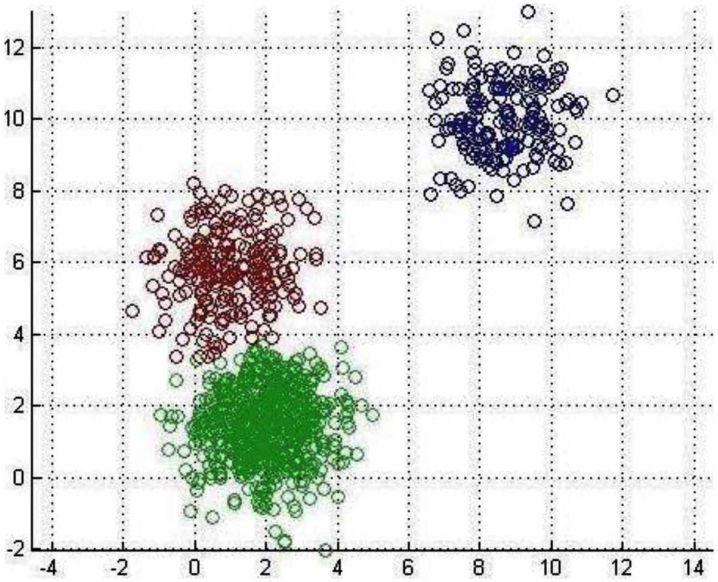
The illustration of the K-Means algorithm execution result will help you understand how it works (Figure 16).

From figure 16 we can recognize three clusters:

Cluster 1: In red
Cluster 2: In blue
Cluster 3: The green one

Data with one to three dimensions can be represented graphically. The others can only be mathematically apprehended. Appendix 2 simplified the representation of the iterative process that the algorithm will perform on two-dimensional data.

Figure 16. Example of k-means clustering



So, this algorithm solving the following problem: "Given points and an integer k, the problem is to divide the points into 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). You're going to think that's pretty much the same thing, but that allows as 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 K-means 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 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. While 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 be very well that our population has characteristics that make 3 clusters partition it better than 2 clusters. This can precede our K-Means Clustering with a *Hierarchical Classification* that will automatically define the best number of clusters to choose from.
- 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 that I did not present here, I proposed a solution to initialize fictitious centers of gravity on a line crossing all the dimensions of the data.

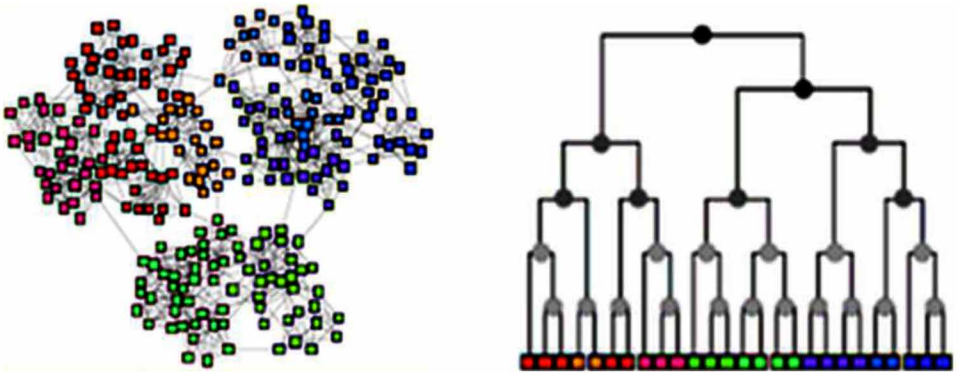
Hierarchical Classification

The growth of data available today generates new issues for which statistical learning does not have adequate answers. Thus the classical framework of classification which consists of assigning one or more classes to an instance is extended to problems with thousands or even millions of different classes. With these problems come new lines of research as the reduction of classification complexity that is usually linear depending on the number of classes of the problem, which need a solution when the class number becomes too large.

Among these solutions, we find the hierarchical model. This type of algorithm aims to hierarchically organize the decision functions in a tree structure. The first step of this algorithm is to define a table of distances or dissimilarities between individuals, which will be recalculated at each step. The hierarchical classification will start from the classification given by the singletons; group together the elements (singletons or classes) closest until obtaining a single class.

This method makes it possible to obtain a graphical representation similar to a tree, called a 'dendrogram'. It represents the set of classes created during the algorithm. These have been grouped together as iterations are based on the comparison of distances or dissimilarities between individuals and/or

Figure 17. Comparison between clustering and hierarchical classification



classes. The size of the branches of the dendrogram is proportional to this measurement between the grouped objects.

Graphically, it is therefore at the place where the branches of the dendrogram are the highest that we must "cut", that is to say stopping the distribution into classes. The plot of the evolution of intra-class inertia as a function of the number of clusters also aids the analyst in the choice of K graphically. This decreases as the number of classes increases. Indeed, they are more and more homogenous, smaller and individuals become more and more close to the center of gravity of their cluster.

Since we try to minimize this criterion, we tend at first to select a very high number of K classes. Such a choice would not be wise because the classification would provide very little information. Its interpretation becomes more difficult. This is why it is, in fact, a brutal reduction of intra-class inertia as a function of the number of clusters to be searched.

Different strategies of aggregation between two classes A and B can be considered during the construction of the dendrogram, the best known of which are:

$$\frac{w_A w_B}{w_A + w_B} d\left(g_A g_B\right)$$

- **The Simple Linkage:** Corresponding to the smallest distance among all the distances calculated between an individual of A and one of B;
- **The Complete Linkage:** Corresponding to the greatest distance among all distances calculated between an individual of A and one of B,

- **The Average Linkage:** Corresponding to the average of the distances between the individuals of A and those of B,
- **Ward's Linkage:** Worth considering d as the Euclidean distance and g_A and g_B the centroids of the classes considered, w_A and w_B being the respective weights of classes A and B.

The main question which arises in practice can be formulated as follow: which linkage we must choose and why? To answer this question, general elements lies in the fact that the simple linkage will be associated with the minimal tree. It generates classes of very different diameters and has a tendency to chaining effect: it generally favors aggregation rather than the creation of new clusters and presents sensitivity to noisy individuals.

The complete linkage will create compact classes that are arbitrarily close. It is also sensitive to noise. The average linkage corresponds to a good compromise between separation and diameter of classes; it is, in fact, a good compromise between simple linkage and complete linkage. Nevertheless, clusters will generally have close variances.

Finally, at each iteration (or grouping step), Ward's method chooses to minimize the increase in intra-class inertia. The linkage will tend to create rather spherical classes and equal numbers for the same level of the dendrogram. It also has a certain ease of use: even if the distance between individuals is not Euclidean, its expression is not modified.

It is for all of these reasons that the golden palm of the most popular choice of users undoubtedly returns to Ward's linkage.

It is important to note that the unsupervised hierarchical classification is quite time-consuming. It only applies to datasets that are not very consistent in terms of the number of individuals. The dendrogram is also less and less readable when this number increases.

It should be noticed before closing this chapter that, there are many other methods and algorithms which are developed for massive datasets analysis, and which you need to take into account in order to transform data into useful knowledge (Sedkaoui, 2018b). You should know that sounds statistical methods that are scalable computationally to massive datasets have been proposed, such as:

The Big Data Bootstrap

Traditionally, subsampling has been used to refer to "m-out-of-n" bootstrap, whose primary motivation is to make approximate inference owing to the difficulty

or intractability in deriving analytical expressions (Efron, 1979; Jackknife, 1989). In the massive dataset, there is a serious problem: each bootstrap resample is itself massive. However, in settings involving large datasets, the computation of bootstrap-based quantities can be prohibitively demanding. A new procedure which incorporates features of both the bootstrap and subsampling is known as the Bag of Little Bootstraps (BLB), to obtain a robust, computationally efficient means of assessing estimator quality. This method, proposed by Kleiner and al. (2014), is a combination of subsampling (Politis and al., 1999), the m -out-of- n bootstrap (Bickel and al., 1997), and the bootstrap to achieve computational efficiency. The development of BLB was motivated by the computational imperative; it can be viewed as a novel statistical procedure to be compared to the bootstrap and subsampling according to more classical criteria.

Parallel MCMC

The computational intensity and sequential nature of estimation techniques for Bayesian methods in statistics and machine learning, combined with their increasing applications for big data analytics, necessitate both the identification of potential opportunities to parallelize techniques such as Monte Carlo Markov Chain (MCMC) sampling. In the Bayesian framework, it is natural to partition the data into k subsets and run parallel MCMC on each one of them. The prior distribution for each subset is often obtained by taking a power 1/k of the prior distribution for whole data in order to preserve the total amount of prior information. Neiswanger et al. (2013) proposed to use kernel density estimators of the posterior density for each data subset, and estimate the full data posterior by multiplying the subset posterior densities together. This method is asymptotically exact in the sense of being converging in the number of MCMC iterations. Wang et al. (2015) replaced the kernel estimator of Neiswanger et al. with a random partition tree histogram, which uses the same block partition across all terms in the product representation of the posterior to control the number of terms in the approximation such that it does not explode with m . Scott et al. (2013) proposed a consensus Monte Carlo algorithm, which produces the approximated full data posterior using weighted averages over the subset MCMC samples.

Massive Time-Series Datasets

The detection and analysis of events within massive collections of time-series have become an extremely important task, especially with the development of IoT where data flows come from interconnected objects in real-time.

In particular, many scientific investigations, for example, the analysis of microlensing and other transients, begin with the detection of events in irregularly-sampled series with both: non-linear trends and non-Gaussian noise (Sedkaoui, 2018b). This approach harnesses the power of Bayesian modeling while maintaining much of the speed and scalability of more ad-hoc machine learning approaches.

CONCLUSION

With today's objects more and more connected, data collection is global and massive. What is called big data aims to know and evolve our lifestyles, our uses, but also the way we consume. The ability to process and analyze data is, therefore, a major issue. And this is where the tools doped with algorithms of machine learning will be particularly useful. That leads to a forced wedding between 'big data' and 'algorithms'.

In this chapter, you have acquired the necessary knowledge to develop an effective roadmap for implementing a data analytics algorithm. You have learned how to transform your business objectives into a data analysis process using the data analytics process based on several technical methods. You have also discovered different techniques for advanced supervised and unsupervised algorithms, such as clustering, classifications, and regression models. In another word, this chapter has presented the foundations in order to allow you to grasp the global view, the famous "big picture", which will help you choose the best algorithms.

REFERENCES

Bickel, P., Gotze, F., & Van Zwet, W. (1997). Resampling fewer than n observations: Gains, losses and remedies for losses. *Statistica Sinica*, *7*, 1–31.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Bort, J. (2016, December 7). *How IBM Watson saved the life of a woman dying from cancer, exec says*. Retrieved from http://uk.businessinsider.com/how-ibm-watson-helped-cure-a-womans-cancer-2016-12?r=US&IR=T

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*(1), 1–26. doi:10.1214/aos/1176344552

Kleiner, A., Talwalkar, A., Sarkar, P., & Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *76*(4), 795–816. doi:10.1111/rssb.12050

Krivko, M. (2010). A hybrid model for plastic card fraud detection systems. *Expert Systems with Applications*, *37*(8), 6070–6076. doi:10.1016/j.eswa.2010.02.119

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning (Adaptive Computation and Machine Learning Series)*. MIT Press.

Neiswanger, W., Wang, C., & Xing, E. (2013). *Asymptotically exact, embarrassingly parallel MCMC.* arXiv preprint arXiv: 1311-4780

Politis, D. N., Romano, J. P., & Wolf, M. (1999). Subsampling. New York: Springer.

Quinlan, J. R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies*, *27*(3), 221–234. doi:10.1016/S0020-7373(87)80053-6

Scott, S.L., Blocker, A.W., Bonassi, F.V., Chipman, H., George, E., & McCulloch, R. (2013). Bayes and Big Data: The Consensus Monte Carlo Algorithm. *EFaBBayes 250 conference, 16.*

Sedkaoui, S. (2018a). *Data analytics and big data*. London: ISTE. doi:10.1002/9781119528043

Sedkaoui, S. (2018b). Statistical and Computational Needs for Big Data Challenges. In A. Al Mazari (Ed.), *Big Data Analytics in HIV/AIDS Research* (pp. 21–53). Hershey, PA: IGI Global; doi:10.4018/978-1-5225-3203-3.ch002

Sugiyama, M. (2015). *Introduction to Statistical Machine Learning*. Morgan Kaufmann.

Wang, X., Guo, F., Heller, K. A., & Dunson, D. B. (2015). *Parallelizing MCMC with Random Partition Trees.* arXiv preprint arXiv: 1506-03164.

Watt, J., Borhani, R., & Katsaggelos, A. (2016). *Machine Learning Refined: Foundations, Algorithms, and Applications*. Cambridge University Press. doi:10.1017/CBO9781316402276

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. P. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Wu, C.F.J. (1989). Bootstrap and other resampling methods in regression analysis. *Ann Stat*, *14*, 1261–1295.

KEY TERMS AND DEFINITIONS

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.

Analytics: 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 mounds of internally generated and externally available data.

Big Data: A generic term that designates the massive volume of data that is generated by the increasing 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."

Classification: In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

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.

Deep Learning: Also known as deep structured learning or hierarchical learning is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms.

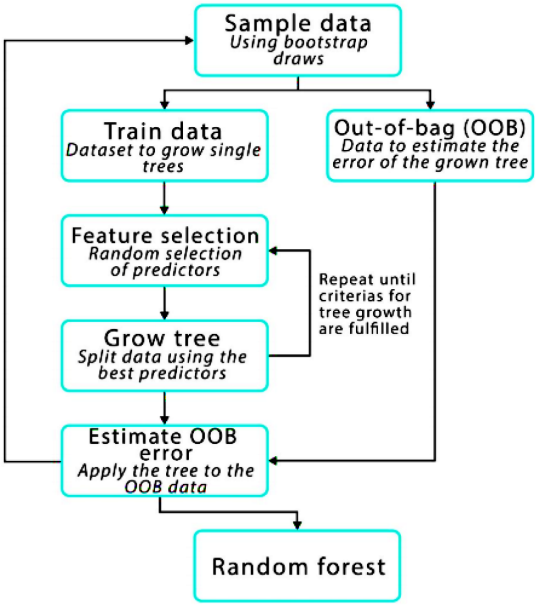
Regression: Regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables when the focus is on the relationship between a dependent variable and one or more independent variables (predictors).

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.

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.

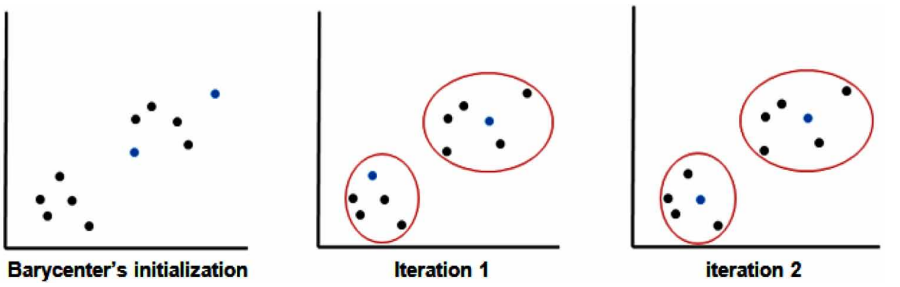
APPENDIX 1

Figure 18. Random Forest Scheme



APPENDIX 2

Figure 19. The Iterative Process



Section 3

Entrepreneur!
Welcome to Your Data-Driven Universe

Being data-driven is a question of actions.
It is not only about looking at the data but also about acting through this data.

By reading the previous sections, you have probably understood how to work with big data and how to apply different algorithms. It is not difficult, and you are certainly feeling ready to make your big data project a reality. But above all, you need to understand your entrepreneurial universe (Chapter 7) to avoid pitfalls and better guide your data-driven business model. Thus, it will be necessary to proceed by phasing and building what is called in the jargon of project management a "road map" (Chapter 8). To better master this domain and build your data-driven culture, this section allows you (Chapter 9) to go to battle through practical examples and applications using data analysis software and technology big data. This section encompasses both theoretical and practical perspectives of big data analytics and its applications in an entrepreneurial context. The chapters in this section discuss different procedures for conducting a particular big data project. Thus, it is expected to act as a practical guide for the students, researchers, practitioners, and especially, entrepreneurs to understand the "knowhow" of different techniques and procedures when working with data. It is hoped to thus enable them to witness a special moment when emerging technologies start to be intensively applied in business contexts.

Chapter 7
Entrepreneurship and Big Data

ABSTRACT

The range of possibilities opened up by big data technologies offers companies in all sectors a remarkable opportunity for development and transformation. And if the majorities are convinced of its strategic interest, many are wondering about the implementation of such a project. Today, companies using big data are search engines like Google; social networks like Facebook, Twitter, or LinkedIn; e-commerce websites like eBay, Ali Baba, or Amazon, etc. But, it would be premature to conclude that big data is reserved for large companies only and that they alone can gain added value from its use. Indeed, as the motorist uses the highway without having built it, the commercial or public organizations, whatever their size and their field of interests, will be able to benefit from the use of big data.

INTRODUCTION

I did it for the buzz. I did it for the pure joy of the thing. And if you can do it for the joy, you can do it forever (Stephen King)

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

It is everywhere, the big data or how the data available today, will change our daily lives. Big data refers to all the data generated every day by a

DOI: 10.4018/978-1-5225-7609-9.ch007

simple contact with our smart object. These data are collected, then stored and especially analyzed! This can then lead to interpretations and strategic decisions for companies. If some people consider big data as a big brother, others see it as a new revolution: that of information!

In what follows, I want to highlight the importance of the big data context for the entrepreneurs or future entrepreneurs who want to start their project in this field. So, listen to me: you have to know that there is the sun for you too in the big data universe. Because to succeed in the digital transition of your company, ensure its development, improve the well-being and efficiency of its employees, beautify its customer relations and fertilizer sales, join this entire universe which born in the heart of the digital transition! "Data is everywhere".

Do you want to accelerate the digitalization of your business? So immerse yourself in your data-driven universe and change your way of doing things. Because this chapter will help you to understand your entrepreneurship context and what is changed in your universe to better prepare your fly to the big data universe and how to join the analytics arena. So, let's go!

ENTREPRENEURSHIP AND ENTREPRENEURS THROUGHOUT HISTORY

Generally speaking, people would be more interested in knowing how the process of entrepreneurship initiates or what the activities of an entrepreneur are (Sedkaoui, 2018c).

First of all, we must qualify the popular belief that attributes the origin of entrepreneurship to economic science alone. A careful reading of the first two authors generally identified as the pioneers of the field, Cantillon (1755) and J.B. Say (1803, 1815, 1816 and 1839), makes us discover authors who were interested in both the economy and businesses, their creation, their development and their management.

Cantillon was basically a banker we would call today lender venture capital. His writings reveal a man looking for business opportunities, preoccupied with clever and economical management that maximizes ROI.

The term "entrepreneurship" can be traced back to as early as the Middle Ages when the 'entrepreneur' was simply someone who carried out tasks, such as buildings and construction projects by applying all the resources that

he disposes of. However, it was during the 16th century when business was used as a common term, and the entrepreneur came into focus as a person who is responsible for undertaking a business venture.

It can be seen that the term of entrepreneur has acquired its present meaning in the course of the 17th century. Even if the term was used before Cantillon, one can notice, as Schumpeter (1954) has noted, that Cantillon was the first to present a clear conception of the whole function of the entrepreneur. So, the notion of entrepreneurship was introduced to political economy in Cantillon's "Essai sur la Nature du Commerce en Général", published posthumously in 1755.

J.B. Say is the second author who has been very interested in the activities of the entrepreneur. He saw the development of the economy through the creation of enterprises. Cantillon and Say saw the entrepreneur primarily as a risk taker since he was investing his own money. For Cantillon, the entrepreneur buys a raw material - often a product of agriculture - at a certain price to transform it and sell it at an uncertain price. It is, therefore, someone who knows how to seize an opportunity to make a profit, but who must bear the risks.

Say will make a difference between the entrepreneur and the capitalist, between the profits of one and the other. In this sense, he associates the entrepreneur with innovation; he sees the entrepreneur as an agent of change. He is the first to have defined all the parameters of what the entrepreneur does in the sense we understand it today.

Schumpeter (1954) himself observed that much of his contribution has been to make the Anglo-Saxons aware of the entrepreneurial world from the writings of J.B. Say. But it is Schumpeter that takes a flight to the field of entrepreneurship and he clearly associated it to the innovation.

Therefore, it was during the 17th and 18th century's Industrial Revolution that business itself was becoming part of the new lifestyle, especially in Europe, where most of this development was taking place. Economists, such as J.B. Say, J. S. Mill, and A. Marshall all included entrepreneurship into the economic spectrum of the time by defining the various skills and features of an entrepreneur.

These definitions vary from an entrepreneur being responsible for employing resources in high productivity areas to earn profits, to risk bearing, and finally to an entrepreneur being responsible for organization and control. However, the most substantial research into entrepreneurial theory was achieved in the 20th century, under the aegis of Schumpeter, who claims that the entrepreneur

has a 'creative destruction innovation' by replacing destroying an existing economy with a better, advances one.

Schumpeter claimed that Cantillon's concept of the entrepreneur followed scholastic doctrine by emphasizing the role of "risk-bearing" directors of production, contrasting with the safety of salaried employees (Schumpeter, 1954). He has not only associated the entrepreneur with innovation, but his impressive body of work highlights the importance of the entrepreneur's role in the development of the economy.

In fact, it is not the only one to associate entrepreneurship with innovation. Clark (1899) had clearly done so before him, Higgins (1959), Baumol (1968), Leibenstein (1978) and most economists who are interested in entrepreneurship. What economists are looking for is first and foremost a better understanding of the role that entrepreneurs play in the economic system.

Since the seminal work of Schumpeter (Schumpeter, 1934), entrepreneurship has been regarded as a positive driving force for regional economic growth and development (Birch, 1987; Storey & Johnson, 1987; Reynolds, 1987; Acs & Armington, 2004). Low and MacMillan (1988) emphasized that entrepreneurship is a process that can be undertaken in a variety of contexts. From this point of view, many studies have indeed stressed that contextual conditions such as education, culture, social support systems, technology, and the presence of human capital and expertise play an important role in the changing conditions for entrepreneurship (Fischer & Nijkamp, 2009).

From what has been just discussed, we can define entrepreneurship as" (Churchill, 1992):

... the process of uncovering and developing an opportunity to create value through innovation and seizing that opportunity without regard to either resource (human and capital) or the location of the entrepreneur – in a new or existing company.

Entrepreneurship has been quite relatively researched and the likely cause could be the generous contributions it makes to public policy goals, such as economic growth, a creation of employment, innovation in technology, enhancement of productivity and structural realignments (Shane 1996). It has a key role in today's economy, is directly linked to innovation, job creation, and the small business sector. It is considered one of the critical elements for a healthy productive economy in developing countries.

In contrast, the research on entrepreneurship started much earlier and traces its roots to different motivations and theoretical concerns. The historical study of entrepreneurship has been particularly concerned with understanding the process of structural change and development within economies. Business historians have focused on understanding the underlying character and causes of the historical transformation of businesses, industries, and economies.

According to P. Drucker (1970) and K. Knight (1967), Entrepreneurship is about taking a risk:

1. It is the process of creating new values that did not previously exist.
2. It is the practice of starting a new organization, especially new businesses.
3. It involves the creation of new wealth through the implementation of new concepts.

Since the 1980s, entrepreneurship has emerged as a topic of growing interest among management scholars and social scientists. The subject has grown in legitimacy, particularly in business schools (Cooper 2003). This scholarly interest has been spurred by a set of recent developments in the United States: the vitality of start-up firms in high technology industries, the expansion of venture capital financing, and the successes of regional clusters, notably Silicon Valley.

Entrepreneurship is emerging around the world as a fundamental pillar of development business' activities. It has a key role in today's economy, is directly linked to innovation, job creation, and the small business sector. It is considered one of the critical elements for a healthy productive economy in developing countries. The entrepreneurial literature has attracted much attention, especially in policy circles (Alvedalen & Boschma, 2017).

Entrepreneurship makes economies more competitive and innovative and is crucial to achieving the objectives of several sectorial policies. Successful entrepreneurship needs, therefore, a dynamic environment.

Maybe you say that this is commonplace nowadays to evoke the constant evolution of entrepreneurial practices. But behind this widely shared reality lie a variety of situations that express themselves in different and complex ways. By their evolutionary nature, it is very difficult to grasp these practices and all the more important to try to do so, because the entrepreneurial practices of today are different from those of yesterday and also from those of tomorrow.

Drucker believes that what entrepreneurs have in common is not personality traits but a commitment to innovation. For innovation, the entrepreneur

must have not only talent, ingenuity, and knowledge but he must also be hardworking, focused and purposeful.

So, they may over time learn to adapt and base the approach they apply to the context of their situation (Haynie & Shepherd, 2009). McGrath and MacMillan identified five characteristics of entrepreneurs (Soyibo, 2006):

- They Passionately Seek New Opportunities: Are alert, always seeking for the chance to profit from change and disruption in the way business is done.
- They Pursue Opportunities With Enormous Discipline: Are not only alert to spot opportunities, but make sure they act on them. Most maintain an inventory of unexploited opportunities and invest only if the competitive arena is attractive and the opportunity is ripe.
- They Pursue Only the Very Best Opportunities and Avoid Chasing After Every Option: Are ruthlessly disciplined about limiting the number of projects they pursue and go after a tightly controlled portfolio of opportunities in different stages of development.
- They Focus Specifically on Adaptive Execution: Rather than analyzing new ideas to death, people with entrepreneurial mindset execute. Yet they are adaptive – able to change direction as the real opportunity, and the best way to exploit it evolves.
- They Engage the Energies of Everyone in Their Domain: Involve many people, inside and outside the organization in the pursuit of an opportunity.

But, today's entrepreneur is fundamentally different in the form of yesterday's entrepreneur. The evolution of new technologies is for many as well as access to quality software solutions at a cost now not prohibitive. Indeed, in recent years have been marked by the massive decentralization to advanced methods and IT tools formerly available only by large groups. This gives access to ever larger numbers the means to be creative and innovative.

The entrepreneurs of 2018 are organized differently from those of 2008, who, in turns, worked differently from 1998 entrepreneurs. With ever more mobility, offices are decentralized, teams working remotely, using new technologies to share their work. The individual is now at the center of new businessdevelopment,managementishorizontal,moreresponsibleandmore empowering. Today's entrepreneurs are built around projects that are often passionate and always ambitious. They are more than anything concerned by

the results of their business and by the incredible number of opportunities that remain to be seized!

DIGITAL REVOLUTION RENEW THE VIEWS ON ENTREPRENEURIAL PRACTICES

Historians will remember from the beginning of the 21st century a profound change in the economy, the "digital revolution", where the dominant economic actors are no longer the actors from industry, energy or physical distribution, but those from digital universe, the famous GAFA (Google, Apple, Facebook, Amazon) in the West, and BAT (Baidu, Alibaba, Tencent) in the East.

This revolution is of the same magnitude as the industrial revolutions of the 19th century or that of printing at the beginning of the Renaissance but on an accelerated timescale. This is just the beginning. All companies are equipped with e-mails, web, mobiles, and IS (ERP, CRM ...) which have become standards: this is no longer the issue.

All advertisers have realized that no marketing campaign can imagine without a digital dimension and that the paradigm is reversed: many advertisers are just digital marketing. This has profoundly changed the customer relationship.

All retailers have realized that consumers often learn and compare very actively on the internet before buying. They also understand that facing giants like Amazon, there is no choice but to excel in the customer experience. Digital has also become the main vector of many business activities in many sectors such as health, recruitment … The famous IoT is also creating significant productivity gains.

You have to remember that in the 1980s there was a lot of talk about Silicon Valley. This is where the important things were happening in the IT field. IBM (mainframe), DEC (mini-computers), Texas Instrument and HP (pocket calculators) dominated their rivals. And then, a new wave of innovation has arrived, Intel Microprocessor, IBM PC, Microsoft MS-DOS, and of course HP Laser Printer.

Digital offers opportunities to transform a sector or an industry through a value chain vision into a much larger digital ecosystem. This paradigm shift involves new modes of interaction between the players in this digital ecosystem. This leads to rethink the foundations of traditional channels and to decipher new forms of collaboration.

Thus, this revolution is also an "Entrepreneurial Revolution". Millions of young and old entrepreneurs launch their startups with very low costs and huge ambitions. They have their Bible "The Lean Start-Up", with its proven methodology: identifying a problem to be solved on a large target market, sourcing user needs and insights, prototyping, testing, and launching a minimum viable product then scaler it as quickly as possible. They go very fast because they have resources for a limited time, and they are often de facto subject to a race with their competitors.

Some succeed by bursting value chains and strongly jostling established actors: Uber for taxis, BlaBlaCar for intercity transport, Airbnb and Booking for hotels, Netflix for TV ... and this is of course only the beginning because the trend is accelerating and industrializing.

In recent decades, entrepreneurs have managed to transform a scientific discovery in the field of computer science into a cluster of innovations. Some of them are visionaries, passionate about research and evolution, and have left their traces by distinguishing their business model in the history of entrepreneurship in the digital age, such as:

The Duo That Revolutionized the World!

Larry Page and Sergey Brin are two entrepreneurs who have revolutionized Internet research. They develop upon their activity from Stanford University a large project called Google. It offers a new, smoother and more convenient dimension to Internet searches. They create online search software based on a predefined algorithm. Google was a real success in 1999. In two years, the requests to be processed increased from 100.000 to 100 million per day. Four years after its creation, Google has achieved unprecedented financial success, thanks to advertising inserted on the platform in a subtle and space-saving way. Today, it is the most used search engine on the web.

Bill Gates, the King of Computing

Bill Gates, the founder of Microsoft is also among the entrepreneurs who left their traces in history. He started from the belief that every home can have access to computers. In 1975, he and Paul Allen, a pioneer in microcomputer technology, created a program for Altair computers. This project brings them together $ 3000. A year later, the two "geeks" leave the university and decide

to found Microsoft. They primarily base the MS-DOS operating system for IBM computers and charge a commission on each PC sold. By owning the exclusivity of their operating system, the two partners then managed to equip computers around the world with their own software.

Jeff Bezos, the Giant of E-Commerce

Jeff Bezos, the big boss of Amazon has left its mark on the world of e-commerce. This passionate physicist, and computer scientist, has been able to draw the vein of the Internet, to set up an online trading platform and offer products that bring a plus to the classic distribution model. The idea came to his mind as he read the rate of growth of the Internet at 2300% per year. To stand out from the competition, it offers in addition to these flagship products, a faster delivery time with a follow-up and reliable after-sales service. It even offers its customers products similar to what they ordered.

And Jack Ma

There is also Jack Ma, the founder of Alibaba.com, a website that allows Chinese entrepreneurs to connect with foreign companies to become their subcontractors. The website has met with tremendous success, especially after the 2003 SARS outbreak, which limited the movement of Chinese businessmen abroad. All transactions are therefore remote, thanks to Alibaba.com.

Ultimately, successful entrepreneurs will either use a technological axis, or use existing technology in a new field, or look for a differentiation axis based on the observation of the expectations of a target thanks to the famous "design thinking".

Practitioners discover innovative strategic models that break with existing business models. By joining the digital world the entrepreneurs are able to:

- Develop Their Digital Culture: Understand digital technologies and measure their implications in their business environment;
- Change Their Strategic Perspective: Experiment with new business and managerial practices inherent to digital transformations;
- Challenger Their Business Model: Modify perceptions and formulate a corporate strategy in line with the new rules of the game of their digital ecosystem;

- Identify and Support Innovation: Master the sources of disruption and the impacts of these innovations to anticipate and seize business opportunities;
- Evolving the Organization: Reinforce and enhance their acting position change and mobilize their teams by placing collective intelligence at the heart of interaction systems.

For the founders of Uber, or Netflix ... success goes hand in hand with the ability to be agile, it is even the key to success! Faced with the digital revolution that no business can escape, the entrepreneur is in the front line. He must transform the organization of his business activity, but also rethink its own mode of management.

For Schumpeter (1954), greed is not only what motivates the entrepreneur. It is primarily animated by an adventurer and seeks the sensation of conquest and discovery. The innovations they create change in the lives of thousands of people. Profit is legitimate because it serves to reward the risk-taking of the entrepreneur. For him, the entrepreneur in a competitive market is the engine of growth.

His theory was useful in his time to understand the changes that accompanied the Industrial Revolution. It is equally so today to understand the transformations that accompany the digital revolution taking place nowadays.

The digital revolution is on the move. Simply put, it consists of digitizing all the activities of the economy and constantly connecting all the elements involved in the production and sale of these goods and services. Worn simultaneously by different elements such as big data, the IoT or machine learning, the digital wave fascinates as much as it worries. Everyone is talking about it, but we are still struggling to see what it really looks like.

To advance in the digital revolution and make relevant connections with the reality of their business, an entrepreneur needs concrete benchmarks and keys to reading. He must take advantage of the specific strengths. Flat hierarchies and more agile structures encourage transversal projects such as big data projects. It is also a way to strengthen the corporate culture by proposing an innovative and integrative idea.

For this, it is necessary as an entrepreneur that you develop your inner strength, show vision, courage, self-denial, pugnacity, and indifference to criticism, pressure, lobbies ... you will discover that the way is more important than the goal. In order to cope, you will display more attitude than aptitude, more convictions than certainty, more skill than technique. Because it is your posture of opening to the ideas and of the desire to change that will make the difference. In cohesion, you have to relay, explain, train...

The flip side of the coin is a misuse of the big data with the look on the almighty state "Big Brother is watching you" mentioned by George Orwell in his book 1984. The risks exist and they are very real but if you want to move forward, the digital revolution is a mandatory passage, it is up to you to be the actors and not the spectators. So, unleash the entrepreneur who is in you and try to develop your big data project, because, the digital age forces you to reinvent values and approaches in all areas.

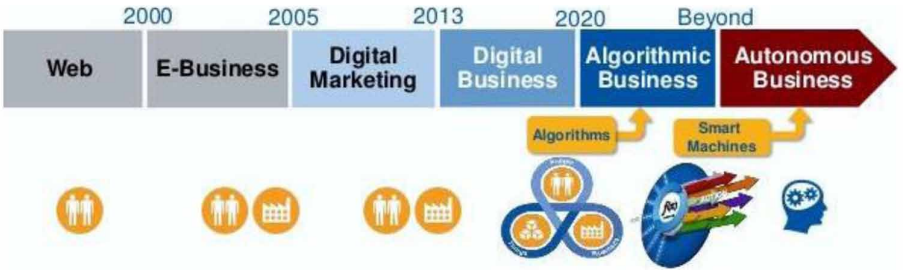
WHAT KINDS OF ISSUES WILL ENTREPRENEURS DEAL WITH IN THE BIG DATA ERA?

Entrepreneurship is one of the most dynamic approaches to socio-economic transformation and development. It is intimately related to private sector development, micro, small, and medium-sized enterprise policies, job creation, innovation, and competitiveness. Different nations understand entrepreneurship as an indispensable element to preserve the viability and competitiveness of a country's economy.

Entrepreneurship is considered as the study of sources of opportunities; the process of the discovery, evaluation, and exploitation of opportunities; and the set of individuals, who discover, evaluate and exploit them (Shane and Venkataraman, 2000). Entrepreneurship nowadays plays a vital role in creating opportunities for individuals and in their ability to meet future challenges.

If we look on the history of our society, we will see that it is characterized by the stages of evolutionary change such as "Agricultural Age", "Industrial Age", "Information Age", and "Technology or Knowledge Age". At each stage of the development, people have existed, lived and worked together in

Figure 1. Business changes with digital revolution
Source: The Gartner Group, 2016



order to advance the level of technology, improve living conditions, increase the economy, etc.

The industrial revolution in the 19th century and digital revolution in the 21st century: 200 years apart and yet with striking similarities. The common point of these two periods lies in: the development of the means of communication. The invention of the telegraph and the telephone for the first, and the development of the Internet and IT tools for the second, so many innovations which allowed a strong business development (see Figure 1).

During each stage of evolution, entrepreneurs are facing many challenges (Sedkaoui, 2018c). So, they must develop, to adapt to each stage, a new entrepreneurial culture, adjusted to society and based on knowledge, innovation, and involvement of new methods and strategies in entrepreneurial projects.

To enable the development of the entrepreneurship field entrepreneur agenda needs comprehensive adjustments, and refocusing on new areas and other factors shaping entrepreneurial ecosystems, such as big data analytics, data science, machine learning, etc.

With small budgets, limited staff and inexperience, they somehow have to find a way to compete against many large companies. Nevertheless, there is a school of thought which says that (Stevenson & Jarillo, 1990; Stokes & Wilson, 2010):

Being entrepreneurial means that an opportunity must be pursued despite the lack of resources, and the ability to leverage external resources is one of the hallmarks of the entrepreneur.

In this term, it's necessary to understand the leadership's cognitive preference in order to determine which encourage the adoption of big data analytics, as suggested by McAfee and Brynjolfsson (2011) and Ross et al (2013).

Big data is opening up a number of new areas for entrepreneurship and investment. Products that make data more accessible, that allow analysis and insight development without requiring the entrepreneur to be a statistician, engineer or data analyst are one major opportunity area (Feinleib, 2014).

Just as Facebook has made it easier to share photos, new analytics products will make it far easier not just to run analysis but also to share the results with others and learn from such collaborations. There is certainly a huge career opportunity that came from being an adopter and integrator of new

analytics tools and technologies, data mining, predictive analytics and data science included.

Actually, with the increasing integration of different technologies in a growing range of equipment and products, entrepreneurs and startups operate within a changed environment. With the developing of the IoT and the coming of the semantic web, new methods for representing, storing and sharing information are going to replace the traditional systems. Offering to businesses and decision-makers, unprecedented opportunities to tackle much larger and more complex big data challenges.

In addition to ICT's advent and speed production and dissemination of these data and their processing capabilities, another element becomes important in recent years: "time" (Sedkaoui and Monino, 2016). This implies a temporal element concept of information flow speed. This calls to rethink the strategy of entrepreneurs, beyond the difficulties posed by the processing of large amounts of data.

It can be seen that before thinking about operations and tools, this phenomenon generates cultural and managerial transformation. It is a major strategic issue. This is a global project that must be carried by the entrepreneur. Because he is the only one to be able to impulse a movement which, for sure, will heckle violently the established order. The exercise is difficult because it will challenge the powers and influences within the company.

To succeed, it must, therefore, anticipate the resistance to change, and quantify it accordingly in the RIO and avoid being content to rejuvenate old principles. As Keynes said: "The difficulty lies not so much in developing new ideas as in escaping from old ones". It is then about doing otherwise.

Small businesses are struggling to address issues related to the data and how to use them to optimize their performance. Yet, they are innovative and agile businesses that have the potential to be able to seize this issue. Despite their dynamism and optimism, these small structures face specific challenges and greater complexity than larger companies. They face greater international competition. They are in fact forced to have processes that are at the forefront of innovation. And when we know the place of data in the decision-making process and in the business playground today, we must ask this question: "What keep these small structures from tackling these issues, and what is needed to be done?"

To be able to address big data issues and challenges, these small businesses must first be able to combine a set of skills. This included:

- Be able to collect and store data (IT tools)

- Be able to analyze them despite their volume, variety, and velocity (data analytics process)
- Be able to derive the relationships in databases and projected them to address problems (business)
- And to convert the results in a clear and intelligible form in such a way that allow these small structures to be able to act on these conclusions (action)

The problem is that these structures often have difficulty managing this data or even for attracting talent that can combine all this data and turn them into knowledge It is possible to call on external expertise to start developing the analytics competency or to tackle a specific problem that requires specific know-how. But the question of outsourcing, for small businesses and entrepreneurs, deserves an increased intention.

Another important aspect is the one related to the technology. Among the plethora of technologies available solutions, which technologies can these they adopt? Should they choose Software as a Service (SaaS) cloud solution, for example? Or it is better to choose one of the many open source solutions?

The pitfall of these small structures or entrepreneurs is that they are in direct competition with larger companies without having the same tools. The questions related to the costs when working with data and its applications are crucial: analyzing a thousand or a million of data is fundamentally no different from a statistical point of view, but the cost often remains the same? It is necessary therefore to find a solution that guarantees a positive return on investment.

The key idea in data analysis is to work incrementally. For entrepreneurs or small businesses, it is not necessary to proceed directly to the installation of the tools that will integrate the entire chain of value creation. It is quite possible to go first with an open source solution, like the Hadoop, Map Reduce, or SaaS, often cheaper, and once the activity has arrived at maturity in big data context, then the entrepreneur can move to more integrated solutions such as SAS. This kind of software provider also offers flexible solutions that can adapt to the specific needs of its structure.

Another important point is that related to the nature of the data, which can be produced, collected and used by the different services of the company (finance, marketing…) In essence, these data are used to optimize operational or sales processes… which can be considered as a complex project, because it involves many services at the same time.

Getting to board all impacted people often presents a difficult challenge for companies.

But, it is not the case for entrepreneurs and small structures, because this last point is actually an advantage for them. They are characterized by more agile structures and more flat hierarchies than larger companies. In addition, they generally demonstrate a very strong corporate culture and greater cohesion than in organizations where the number of employees makes relationships less personal.

Continuous and rapid changes, the digital revolution, new entrants, innovation, big data, IoT ... are all factors that force companies to continually adapt this changing environment that some people call 'VUCA', or: 'volatile', 'uncertain', 'complex' and 'ambiguous'. To survive in such an environment, entrepreneurs must be agile in terms of supply, market, internal organization, income model, etc. Differentiation with competitors is no longer just about products or services, but about how they create, deliver and capture value, in other words, the differentiation is done on the "Business Model".

No doubt, the data revolution profoundly changes the economic landscape. All businesses are impacted. New actors disrupt every sector of the traditional economy. Known by the acronyms GAFAM, NATU, BAT, FinTech, HealthTech and other FoodTech, they are shaking up the established order by creating new business models.

DATA-DRIVEN BUSINESS MODEL

When one considers the opportunities offered by big data universe, the power of analytics, algorithm relevance and of what may seem to be revealed by each byte of data, and then the effort involved seems to be doubled to start down into how one can develop the new business model through joining big data analytics arena. In another way, every data byte tells a story and data analytics, in particular, the statistical methods coupled with the development of IT tools, piece together that story's reveal the underlying message (Sedkaoui, 2018a).

Many successful entrepreneurs' experiences support that, 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, Travis Kalanick and Garrett Camp of Uber, Reed Hastings of Netflix and more. At this stage, one must wonder 'how they do what they do?' Somehow, the

answer lies in the fact that these experiences have understood the underlying message revealed by the amount volume of data byte available today.

They have seen the potential in using analytics not only to differentiate their business models but also to innovate. And innovating in a business model is about exploring new ideas, testing new value propositions and setting up new value chains.

Since the mid-90s and the advent of Internet startups, the term business model is probably among the most used in the business world even though it has no clear definition. The concept refers to how companies do business, in another word, how they compete and make profits by using their competencies and resources to sell goods and services in the market (Zhu and McKelvey, 2013). Drucker (1994) defined a business model as:

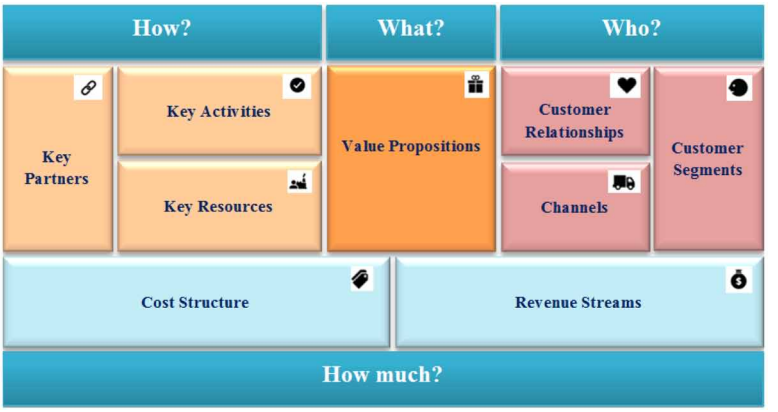
What an organization is paid for, what an organization considers being meaningful results (how to make a difference) and where an organization must excel in ordering to maintain leadership.

The business model is an essential element for a company's strategy but often overlooked. It describes how the business creates, delivers, and capitalizes value and benefits. But it must not be frozen. In a context of widespread digitization, evaluating and evolving your business model is not an option, because everything changes quickly and only the agilest companies will resist. Creating and delivering value to the customer, shareholders, employees and all stakeholders in the organization, these are the main instrument of business success. So, 'value' occupies a central place in the definition of the business model.

To help the entrepreneur in this essential strategic exercise, a tool was developed: "the Business Model Canvas". Ideal for taking stock quickly and visually, this canvas facilitates diagnosis and helps manage complexity.

By its ability to make the apparent complexity of an organization's business model intelligible, attractive and operational, this canvas, which is the result of the Swiss's researcher/entrepreneur work (A. Osterwalder), has quickly and widely established itself in the business world. In just a few years, it has become the absolute reference in the business model. The canvas consists of modeling and simplifying the main elements of an entrepreneurial activity, i.e. the process through which this activity, delivers and captures value. This layout facilitates the description; the definition and the analysis of the interactions of the different parts of the model (see Figure 2).

Figure 2. Business Model Caneva



This modeling facilitates the global vision of the interactions between the bricks constituting the activity and makes it possible to ask the right questions. What is the promise of the offer, what solution does it bring to the customer's expectation, and besides, to what type of client is activity's value proposition addressed, and how will it be done? Etc.

So, to generate value it is essential to ask the following questions:

- What and Who: Your offer or the value you offer and for whom (customers, products or services, price…).
- How: Your resources and skills needed to architect value (value chain).
- How Much: Profitability that combines revenue and costs (employed). It is the financial translation of value.

PwC (2016) discusses five key disruption factors that can change business models:

- Consumers: Or how to respond to more demanding customers, more volatile, better informed and less loyal customers.
- Competitors: Or how to deal with the growing number of young, more efficient and agile businesses that disrupt every aspect of the economy distribution: how to seize the opportunities offered by the multiplication of channels, connected objects, artificial intelligence, big data and the sharing economy.

- Production: Or how to adapt processes and production methods by integrating new technologies to accelerate the development of products and services.
- Regulation: Or how to adapt to new, stricter regulations, to which many sectors are subject (finance, energy, transport, ...)

But, when the entrepreneur starts making more tactical decisions, to develop his business project, data always are helpful. The examples cited throughout this book illustrate this best. If these companies and other were able to surprise us it is because they carried out a series of successive business model innovations oriented towards the 'data' and the 'Analytics'.

Being 'data-driven' means making decisions based on data. Traditionally, decision-makers and CEOs make decisions based on the goal they have set. Data-driven companies focus on collecting and efficiently analyzing data and make decisions on that basis. Look at businesses like Amazon, where even the smallest decisions are data-based, even the color of the walls was decided by data. Whenever they have an idea, they analyze the data to validate it. Data is a key success factor and the companies that make good use of it will gain in competitiveness.

Data-driven business model (DDBM) puts data at the center of value creation. This central place can be translated in different ways: analysis, observation of customer behavior, understanding of customer experience, improvement of existing products and services, strategic decision-making, and marketing of data.

For the last case, data may be the main building block of the company's offer. Thus, we will find in this category companies that offer data, whether financial (Bloomberg, Reuters …), economic (Dun & Bradstreet…), or from the media social networks (Gnip, DataSift, etc.). This data can be aggregated from different sources, generated directly by the company, processed and enriched by various analyzes and highlighted by data access and visualization platforms. As for revenue models, these can be based on a direct sale of data, a license, a subscription or a free provision financed by advertising.

More precisely, among the different big data opportunities, it is possible to identify three broad areas, (business model around big data in some way) in which it creates value and has an impact on companies:

- Data as a Service: This first business model is intended for companies that generate a large amount of data, but do not have the means at their disposal to collect or make them in forms that can be analyzed.

Many public institutions use this model. Municipalities, for example, generate transportation data, and companies can seize it for their own users.

- Information as a Service: In this case, the product provides is directly the information obtained from the analysis of the data. Companies that allow their users to monitor their physical activity, such as the number of steps taken during the day, corresponding to this model. Fitbit users, for example, produce the data and they pay for their visualization in graphical form.
- The Recommendation as a Service: This model is the most lucrative. In this scenario, the product provides is directly a specific recommendation addressed to users of the service to guide their consumption choices. Services like Mint.com offer their users to view their accounts and spending on their different credit cards to get a unified view of their budget. In exchange, Mint allows financial institutions to offer their products in the form of personalized recommendations.

Beyond these well-known business models for which data creates value, there is another category where the data serve as a value creator, notably through the exploitation of big data, without necessarily being present directly in the offer.

In this category we find all companies relying on the use of big data to derive value. For example, we can mention GAFAM or Uber, which uses data related to the location of its clients, its drivers, trips, traffic to determine the price, or Netflix, which uses the collected data about the use of its streaming service to identify themes and concepts of new series, 'House of Cards' is a great example of success.

But, despite this variety of applications of data, to date, there is no global and unified model in the literature to address it as a project and thus define its scope according to the practical needs of the company. An effort of conceptualization in this sense is then necessary. A draft model, which could be proposed, would distinguish these elements influencing such a project:

- The Data: In relation to their nature and thus the three Vs, and other additional Vs described previously, their level of confidentiality of that data, protection of the private life or defense secret, for example, is to be taken into account in a big data project.
- Data Flow: The nature of data flows would also be important. In the information systems, there are three types of flow: decision-based:

with respect to an event; informational: characteristics of a business object such as a bill; and physical: the stock of products or financial transaction.

- Data Sources: Data can come from different sources such as documents, structured databases, log files, social networks, mobile phones, Wi-Fi terminals, cameras, and geolocation tools (GPS), sensors (Sensors), etc.
- The Data Context: That would have its great influence in such projects as for example the type and size of the company, its sector of activity, the process or the trade in question, the duration of the data collection, the urgency of data processing, data security, etc.
- Analytics Methods: According to the preceding parameters, the choice of the algorithm or the method of analysis is necessary. These methods include sentiment analysis, graph analysis, predictive analysis, simulation, pattern recognition, data visualization, segmentation, classification, data mining, genetic algorithms, machine learning, regression, signal processing, etc.
- Technologies: The selection of technologies at all levels, such as applications, software, and platforms, technical architectures ... could be crucial. Among the technologies at the software level, note, for example, Hadoop, MapReduce, HBase, NoSQL, Spark, etc.

Table 1. Important dimensions to boost big data project

Dimension	Objective
Strategic alignment	Including the business's strategy, its business model, the technologies used and the skills mobilized
Value creation	Operational excellence and process optimization, gaining market share, preventing problems and reducing errors, improving customer service, etc.
Performance Measurement	Appraised by indicators and/or metrics at the intrinsic, for example, the customer satisfaction, and extrinsic such as turnover and net results levels
Resource management	Involves their allocation, combination, and optimization
Risk management	Takes into account computer security and data piracy, technology dependence or technology provider, etc.
Management Responsibility	Includes access control and entitlements, process traceability, data privacy …
Management Capacity	Includes employee skills management, f for data analysis, financing capabilities, robust functional and technical infrastructure

In addition, other variables or essential decision criteria that must be taken into account in the implementation of a big data project such as its objective, the issues to be taken into account, the budget, the project timing, available resources, necessary skills, etc.

Specific objectives when working with big data include, for example, event prediction, risk prediction, fraud/crime detection, flow optimization, data qualification, Open Data, the quality of services, and the detection of business opportunities, etc.

Also, a big data project cannot be done without being part of the massive data governance (detailed in chapter five) coupled with the organizational IS. It involves taking into account the dimensions illustrated in Table 1.

Big data analytics has become an essential requisite to run most businesses. Though startups and young entrepreneurs might not spend much on big data but definitely need to research, train and follow the trend to take their ventures to heights and create their own DDBM. Integrating big data analytics can generate many advantages for the entrepreneurs, such as (Sedkaoui, 2018c):

- Supporting Decision: As mentioned before, entrepreneurs 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 the 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 startups and entrepreneurs can be successful in cutting down their operational costs, which is typically one of the biggest challenges for every fresh venture.
- Customer Insights: The growth of any company depends on how to keep the preferences, likes, tastes … of their customers into account to design 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 Use: Over the last year there has been an increase in the perceived use of open data by entrepreneurs to build new products and services. Open data, in addition to its potential economic, and creation of new activities also fall within a philosophical choice or ethics. They encrypt collective human behavior, and therefore also belong to those we measured these behaviors. The culture of this phenomenon builds on the availability of data to a communication orientation.
- The Role of the Entrepreneur is Therefore Central: He alone has the business vision that allows extracting the value from big data to differentiate his business model. It is up to him to define a working strategy of the big data project based on business imperatives.

DEVELOP A DATA-DRIVEN CULTURE TO BETTER ORIENTATE ENTREPRENEURIAL ACTIVITIES

From the reading of the DDBM typology mentioned previously, we can see that big data use is defined taking into consideration, three essential points: (i) its finality, (ii) the skills needed, and (iii) the organization to be adopted. This is to say that, a company carrying out a big data project makes choices in terms of skills to mobilize, organizational transformation to implement depending on the purpose of use of the data it has set.

Big data and analytics tools help entrepreneurs to orientate their entrepreneurial approaches and make the decision easier. Entrepreneurial orientation is defined by Rauch et al, (2009) as:

Policies and Practices That Provide a Basis for Entrepreneurial Decisions and Actions

Entrepreneurial orientation is another strategic orientation that has been linked to firm performance (Lumpkin & Dess, 1996; Wiklund, 1999; Zahra & Covin, 1995). It represents a set of organizational capabilities for innovative firms (Miller, 1983).The most commonly used dimensions of entrepreneurial orientation in the literature are: 'innovativeness', 'risk-taking' and 'proactiveness'.

The two dimensions of entrepreneurial orientation which point to the link between entrepreneur and big data capabilities are:

- Innovativeness: Describes the willingness of the business to introduce novelty technological leadership in developing new processes (Lumpkin and Dess, 2001). Achieving an entrepreneurial mission through innovativeness refers to the ability to solve business problems or in effect to create significant value (Sedkaoui, 2018c). This supports a contention that it will be a key indicator as to whether entrepreneurs, in many sectors, will adopt big data analytics.
- Proactiveness: Is a forward-looking perspective that suggests that companies with high level of entrepreneurial orientation will be looking to be the first to market capture a particular segment and act in advance in anticipation of future demands (Lumpkin and Dess, 2001). Proactive companies can make use of big data analytics to improve their understanding of their customer and their sector, with the condition that they have access to the right sources or 'data' (Sedkaoui, 2018c). It is, therefore, an important element to consider when looking at big data adoption in small and micro enterprises. For example, by retaining the environment, this reflects the tangible and intangible results of breaking patterns, changes in the system, and new discoveries towards process improvement.

The literature suggests that entrepreneurial orientation is a useful lens through which to consider the use of big data analytics in small businesses. And the contention made by McAfee and Brynjolfson (2012) and Ross et al (2013) suggests that a culture of evidence-based decision-making is required for the successful adoption of big data in companies.

There are two ways to transforms data into a valuable contribution to a company (Sedkaoui and Monino, 2016):

- Transforming data into information is one of the stages of data value production, which is exploited in order to obtain useful information and to successfully carry out company strategies. This automatically involves database information in company decision-making processes;
- Transforming data into products or processes adds value to companies. This is produced when data analysis must be implemented in the physical world.

It's clear that the ability of entrepreneurs to adopt big data analytics may be understood by looking at their role in the determination of the entrepreneurial culture and how they are deploying their resources to engage with and make

use of analytics tools and methods in their field. For them, it is essential to have data, increasingly, many on the environment in which it operates or will operate.

In order to succeed in an analytical approach and boost a big data project, it is necessary for an entrepreneur in business and science field, or even in the social field, to prepare it in advance. To do this, three essential questions must be asked: why, what and how.

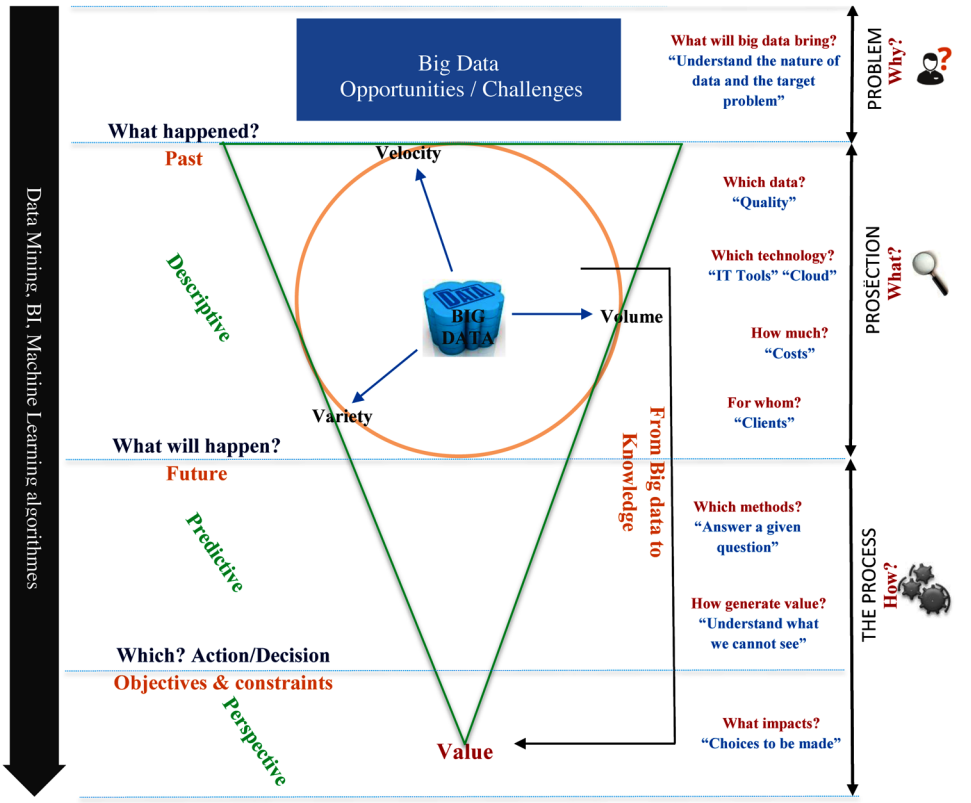
- Why: The first question to ask is "why? ". In most cases, this question will inevitably occur during the initial briefing with a consultant or client. Many big data projects are launched only because the term big data is in vogue. Many executives board the wagon and begin to approve massive investments of time and money to develop a data platform. Most of the time, this strategy is based entirely on the motive that "everyone is doing it". An in-depth analysis of the goal that an entrepreneur wants to achieve, by accessing this data, as well as an assessment of the investments and expertise that the project needs, are required but too often overlooked in the context of the deployment of a big data strategy.
- What: In all sectors, companies are now considering turning the corner on big data and analytics. They recognize in the data a largely untapped source of value creation and an exclusive factor of differentiation. But, many don't know which approach to tackling. What entrepreneur is trying to do? Does the project want to create an innovative market, or find a new channel that requires information on client interest and future profitability? Etc.
- How: While companies do see the great potential that big data analytics can bring to improve their business performance, the reality is that many are struggling to generate value from available data. Gartner (2016) study shows that many big data projects remain blocked and that only 15% have been deployed in production. Examining such failures, it appears that the main factor is in fact not related to the technical dimension, but rather to the processes and human aspects that prove to be as important. Conduct a data-driven project means also to be able, in particular, to answer questions such as: How can we be sure that big data could help us to create business impact? Who should be involved and when? What are the key steps that need to be attentive? Is the project on the right track to succeed? Etc. It is therefore essential, for data-driven orientation, to ensure:

- For the data: Quality, security, structure …;
- For the Process: A well-defined organization, a data-driven culture, its direction …:
- For Tools: IT infrastructure, storage, data visualization capability, performance monitoring.

In order to extract value from big data, it must be processed and analyzed in a timely manner, and the results need to be available in such a way as to be able to effect positive change or influence business decisions. It is important to ensure that the project is progressing towards the intended result (see Figure 3).

Big data is considered a new form of capital in today's marketplace (Mayer-Schönberger & Cukier, 2013a; Satell, 2014), many firms fail to exploit its benefits (Mithas, Lee, Earley, & Murugesan, 2013). For entrepreneurs, it

Figure 3. Orientate entrepreneurial approach based on data-driven



is known that they are unlikely to analyze data on the same scale as large companies like Google, Facebook, Walmart, Twitter, Netflix, Amazon, IBM, and others, due to their limited sources, skills and IT tools.

Yes! It's possible that they are engaging with the free big data tools provided by companies like Google, without forgetting the increase in the prominence of social networks and the fact that engaging with social media which can help generates exposure and traffic for entrepreneurs at a much lower cost than traditional marketing approaches (Schaupp and Belanger, 2014). But, it's to highlight that they are unlikely to have large stores and sophisticated tools to capture, prepare, analyze and manage generated data.

Big data is considered a new form of capital in today's marketplace (Mayer-Schönberger & Cukier, 2013a; Satell, 2014), many firms fail to exploit its benefits (Mithas, Lee, Earley, & Murugesan, 2013).

The application of analytics for small and micro enterprises can be divided into three main categories, namely descriptive, predictive and prescriptive analytics (as illustrated in the Figure above).

Entrepreneurs have to expand their efforts to move their small business from using only traditional BI that addresses descriptive analysis (what happened) to advanced analytics, which complements by answering the "why", "what" and "how". Ultimately, 'data analytics' is inevitable as it can help extract various kinds of knowledge from data.

Clearly, the use of big data analytics and IT tools including clouds will provide numerous opportunities to build entrepreneur approach that will effectively and efficiently cater for the needs of the various entities. Therefore, it is necessary to include enough resources and finance to support the analytics' uses by entrepreneurs.

This investment is essential to reap the full benefits of big data and realize all the envisioned features and capabilities. To help optimize the work and minimize costs of such projects it is recommended to include some of the following activities in the process:

- Understanding first what they can do with big data before they consider adopting it.
- Developing analytics tools to help predict and view possible changes and forecast potential problems. This will help avoid or at least reduce some of the risks involved and also help reduce costs.
- Benefitting from other business experiences to follow successful models and avoid problematic approaches.

- Benefitting from experts and researchers to research new possibilities for more advanced analytics that suite the project idea and objectives.
- Combining big data with open data. This can help to reach better decisions and optimize various functions.

Therefore the efforts should concentrate on creating a roadmap for success that covers several stages (Sedkaoui, 2018c):

1. Set up the entrepreneur's direction, in the data analytics context, by identifying the mission, the vision and strategic and operational objectives.
2. Establish policies, principles, resources and expertise guidelines to control ICTs and big data usage.
3. Evaluate and analyze the current situations and the necessary changes and additions to reach the desired result.
4. Identify priorities and use them to determine the most important components and techniques that would offer the greatest effects with the smallest investment.
5. Realize new opportunities for further development by monitoring current analytics developments and their effects and the arising issues and new requirements.

BREAK THE BIG DATA MYTHS AND BECOME A DATA ENTREPRENEUR

Surprising as it may seem to you, we have always driven businesses based on the data.

So, data is absolutely not something new and what is presented as a revolution today is related to how it is produced and used. By the way, I do not know if you have noticed that recently we say more data than big data.

When we remember that in the mid-80s a PC with 128 KB was considered as a powerful machine, and today the simplest smartphone is infinitely more powerful than the Apollo capsule computer that has conducted the Man on the moon, we can see the progress made in terms of storage capacity and treatment.

So, the importance of the data and its power do not lay so much in the key indicator which is often simple and known, but in the understanding of the context, exogenous factors sometimes very far from the business which condition this indicator. It is like the 'butterfly effect'.

Big data is aimed at all types of companies, with the evolution of the market, large companies do not have the exclusivity of the use and applications of this phenomenon. It is true that data revolutionizes businesses activities but creates opportunities for small businesses and entrepreneurs, by helping them in their governance and operation.

When a company claims to want to become data-driven, I personally interpret it as:

- A desire to display "modern" using the buzzword of the moment, or;
- A desire to reassure the market and investors about the business orientation of digital projects alongside initiatives stamped.

But, I hear a lot about deciding, but not understanding. Big data offers great opportunities but requires significant upstream preparation. So, being a data-driven is not the only question of data but also it is a question of understanding that data and also about action.

Monitoring your performance indicators at all times does not make you a data-driven. Running an A/B test does not make your team data-driven product. Having a team of data scientists does not make you a data-driven company.

As you may have understood from reading the various chapters of this book, the first trait of a data-driven business is to make data available: Availability. Here, I would like to draw your attention to two ways to make your data accessible:

- Reactive approach: train all teams to use all the tools of your data ecosystem (which is unrealistic unless you are in a very small structure) so that they can withdraw by themselves the information they consider relevant.
- Proactive approach: push, at a given frequency, a set of indicators (or KPIs) relevant to each recipient without having to directly access the data sources (CRM, web analysis tool and other bases of data).

Thus, being a data-driven means translating your vision into objectives and knowing how to measure success before you notice it: Action.

Also, as pointed out in the MIT Sloan Management Review, "reducing time-to-insight rather than saving money is the primary driver for their big data business investment".

Time-to-insight is the time between data collection and results. This term is often close to "time-to-market", "time-to-answer", or "time-to-decision".

Together, they explain why the growing desire to become data-driven, embrace digital transformation and invest in big data projects.

Data brings a real change of paradigm and way of thinking, so to guarantee the success of a project, it is necessary to ask the key questions, find the right answers, and make the most of it. In order to decode the profit of big data, entrepreneurs must continue to innovate and offer a broad popularization of the business changes linked to the advent of the company's digitization.

Entrepreneurship is not a new concept by any means, but the way in which entrepreneurs function actually has undergone an astonishing transformation over the last decade.

Entrepreneurial leaders have a reputation for being more agile and better at exploiting niche markets than its larger companies. Furthermore, despite constrained resources, they would seem to be ideally suited to take advantage of the opportunities that big data analytics would help identify.

Entrepreneurial in the age of big data, must rely on varied analytical approaches to thought and action to create and implement solutions that are socially, environmentally, and economically sustainable. Data literacy consists of a strategic element and even more especially with the rise of the IoT, sensors and wearable technologies.

The DDBM is growing as companies understand how to use and leverage the huge amounts of available data. We are talking more about smart data than big data. In this way, we assist in the emergence of the companies able to use, exploit and process data to improve their strategies and better guide the decision-making process. Party of the digital industries, these new business models now, benefit other sectors activity, such as hotels, transport, banking, insurance, etc.

The entrepreneur in the big data universe can be motivated by the love of risk, the search for independence, freedom, surpassing oneself, environmental concerns… These motivations will push him to set up a data-driven business. Thus, having skills in data analysis can be useful for this entrepreneur. Skills in machine learning and the use of different algorithms provide an overview that will allow him to better adjust the offer and then his business model.

The entrepreneur initiates the creation of his project oriented-data, from an idea that it makes live. Its business model must be a focus on IT tools and advanced analytical techniques. This project can be deployed thanks to the various technical and business levers setting up.

At first, the entrepreneur creates his business model. This is to assess the viability of the big data project in the short, medium and long-term.

The development of his project must be done tactically so that this project generates value.

Then, the entrepreneur must define the resources deemed elementary. He sets up the tools that will accompany him in the implementation of his project. Here it should be noted that the first difficulty lies in the prioritization of tasks, and in the investment and time allocated to each phase.

In this context, start a value creation experiment using the available data volume requires a methodology and an iterative and dynamic logic. Throughout the entrepreneurial process, the big data project leader will have to go back and forth between the objective of his project and the obtained results (assisting his business model) to change his offer if necessary. To better integrate the market and make known his project, the entrepreneur can then rely on the various networks that surround him.

As we mentioned it, entrepreneurship has advantages and disadvantages at all levels. The entrepreneur will have to protect himself and propose a development plan that he will adjust according to the various hazards related to his environment. To maximize the chances of success of his big data project, he will have to become familiar with the specific tools required, namely:

- Positioning and SEO tools and analysis tools such as Google Analytics.
- KPI tools (Key Performance Indicators) to better identify prospects: number of clicks, advertising by click, the purchase of the product or service itself ...
- The data / web-mining tools, such as Tanagra Software, OLAP models… which are useful for finding very small parcels of homogeneous customers and which will allow the management of the customer relationship.

Being data-driven is based on the actions that flow from these initiatives.

Being a data-driven entrepreneur means being at the heart of data valuing and intervene at all stages of the data chain: problem definition, data collection, preparation, modeling and solution creation. An entrepreneur made in data must know how to present and prioritize the results to be used. So, excellent communication skills are needed.

If an entrepreneur knows how and where to look for the right data (smart data), if he can do this by analyzing a lot of information, everything is for the best. But, if he gets there, by creating an application directly, it's even better.

It is clear that it is necessary to have a thorough reflection to see the essential elements that constitute the springs of the business activity. New

analytics approach in big data age combines predictive and prescriptive analytics to predict what will happen and how to make it happen. Analytics uses and applications improve the efficiency of the decision-making process and generate value. The constraints facing the exploitation of big data and their transformation into knowledge are related to many issues, already discussed throughout this book, such as its complexity, which is growing with the increase of its quantity its velocity and diversification of its types and sources (Sedkaoui and Gottinger, 2017).

Leveraging leading tools and techniques help to manage and extract relevant data from big data. Advanced analytics can range from historical reporting, through to real-time decision support for organizations based on future predictions.

You have to know, that the future belongs to those who make their relentless approach to tracking data and making adjustments based on their findings. Because data has many things to tell but one must know how to make them talk, by considering algorithms (data analytics techniques) as a recipe, data as ingredients, while the computer (IT tools) is like a mixer that supports a lot of the difficult tasks of an algorithm (Sedkaoui, 2018a). It revolves around data, as the digital revolution is continuing, and gives birth to a new concept, a concept made in data in the entrepreneurial' world: "the data entrepreneur".

Becomes a data entrepreneur means knowing how to succeed where others failed.

Several conducted surveys on big data make it possible to know a little more about the elements that still prevent certain companies from taking the plunge. Among the brakes we find:

- Cost
- Lack of skills
- Lack of visibility on big data opportunities
- Companies did not seek to quantify ROI of big data investments, i.e. investments are not weighted by expected earnings)
- Data collection is limited to traditional channels
- More than 85% of available data is unstructured, and businesses do not know how to process and analyze it.

On the other hand, the most mature companies in the big data exploitation are distinguished by the following criteria:

- An anticipation of strategic issues related to better use of internal and external data;
- The diversity of data collected and collection channels;
- Creation of teams or data experts (data scientist);
- Adoption of new data exploitation technologies;
- Better consideration of the issues related to the data protection and privacy in data analytics process.

CONCLUSION

Michael Porter qualifies the innovation as the key to economic prosperity. In a complex, demanding and unprecedented economic environment, innovation seems to be the key to helping businesses succeed and thrive. Among the innovation facilitators we have big data, often described as a panacea in the field of digital business transformation. Without being a miracle cure, big data is above all a powerful tool for creating new uses and therefore building the business of the future.

While big data is undeniably a pillar of innovation, the real challenge lies in its use to best serve the company's strategy. Start-ups, entrepreneurs or small businesses have an advantage over large companies because they are digital from the outset and more agile to integrate big data into their business model.

Major players like Google, IBM, Cisco or Microsoft have invested for several years in the construction of Datacenter but have also deployed solutions dedicated to data analysis. However, new entrants are looking to take a piece of this cake that is envious. In addition, many companies operating in more traditional sectors will be able to take advantage of the big data revolution.

REFERENCES

Acs, Z. J., & Armington, C. (2004). Employment growth and entrepreneurial activity in cities. Regional Studies, 38(8), 911–927. doi:10.1080/0034340042000280938

Alvedalen, J., & Boschma, R. (2017). A critical review of entrepreneurial ecosystems research: Towards a future research agenda. European Planning Studies Journal, 25(6), 887–903. doi:10.1080/09654313.2017.1299694

Baumol, W. J. (1968). Entrepreneurship in economic theory. The American Economic Review, 58, 64–71.

Birch, D. L. (1987). Job creation in America: How our smallest companies put the most people to work. New York: The Free Press.

Brynjolfsson, E., & McAfee, A. (2011). Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy. Lexington, MA: Digital Frontier Press.

Cantillon, R. (1755). Essai sur la nature du commerce en général. London: Fetcher Gyler.

Churchill, N. C. (1992). Research issues in entrepreneurship. In The State of the Art of Entrepreneurship (pp. 96–579). Boston: PWS-KENT.

Clark, J. B. (1899). The distribution of wealth: a theory of wages. Interest and Profits. New York: MacMillan.

Cooper, A. (2003). Entrepreneurship: The Past, the Present, the Future. In Z. J. Acs & D. Audretsch (Eds.), Handbook of Entrepreneurship Research. Boston: Kluwer.

Cukier, K., & Mayer-Schoenberger, V. (2013b). The Rise of Big Data. Foreign Affairs, 92(3), 28–40.

Cukier, K., & Mayer-Schonberger, V. (2013a). Big Data: A Revolution That Will Transform How We Live, Work and Think. Boston: Houghton Mifflin Harcourt.

Drucker, P. (1970). Entrepreneurship in Business Enterprise. Journal of Business Policy, 1.

Drucker, P. (1994). Innovation and Entrepreneurship: Practice and Principles. London: Heinemann.

Feinleib, D. (2014). Big Data Bootcamp: What Managers Need to Know to Profit from the Big Data revolution. Apress. doi:10.1007/978-1-4842-0040-7

Fischer, M. M., & Nijkamp, P. (2009). Entrepreneurship and regional development. Working Paper: Serie Research Memoranda. VU University Amsterdam, Faculty of Economics, Business Administration and Econometrics.

Gartner. (2016). Investment in big data is up but fewer organizations plan to invest. Available at: https://www.gartner.com/newsroom/id/3466117

Haynie, M., & Shepherd, D. (2009). A measure of adaptive cognition for entrepreneurship research. Entrepreneurship Theory and Practice, 33(3), 695–714. doi:10.1111/j.1540-6520.2009.00322.x

Higgins, B. H. (1959). Economic Development: Principles, Problems, and Policies. New York: Norton.

Knight, K. (1967). A descriptive model of the intra-firm innovation process. The Journal of Business of the University of Chicago, 40.

Leibenstein, H. (1978). General X-Efficiency Theory and Economic Development. London: Oxford University Press.

Low, M., & MacMillan, I. (1988). Entrepreneurship: Past research and future challenges. Journal of Management, 14(2), 139–161. doi:10.1177/014920638801400202

Lumpkin, G.T., & Dess, G. G. (1996). Clarifying the entrepreneurial orientation construct and linking it to performance. Academy of Management Review, 21(1), 135–172. doi:10.5465/amr.1996.9602161568

Lumpkin, G. T., & Dess, G. G. (2001). Linking two dimensions of entrepreneurial orientation to firm performance: The moderating role of environment and industry life cycle. Journal of Business Venturing, 16(5), 429–451. doi:10.1016/S0883-9026(00)00048-3

Miller, D. (1983). The correlates of entrepreneurship in three types of firms. Management Science, 29(7), 770–791. doi:10.1287/mnsc.29.7.770

Mithas, S., Lee, M. R., Earley, S., Murugesan, S., & Djavanshir, R. (2013). Leveraging big data and business analytics. IT Professional, 15(6), 18–20. doi:10.1109/MITP.2013.95

Orwell, G. (1984). Big Brother Is Watching You. Academic Press.

PwC. (2016). Becoming a purpose-led and values-driven organization, Annual Report. Author.

Rauch, A., Wiklund, J., Lumpkin, G. T., & Frese, M. (2009). Entrepreneurial orientation and business performance: An assessement of past research and suggestions for the future. Entrepreneurship Theory and Practice, 33(3), 761–787. doi:10.1111/j.1540-6520.2009.00308.x

Reynolds, P. D. (1987). New firms: Societal contribution versus survival potential. Journal of Business Venturing, 2(3), 231–246. doi:10.1016/0883-9026(87)90011-5

Ross, J. W., Beath, C. M., & Quaardgas, A. (2013). You may not need big data after all. Harvard Business Review, 91(12), 90–98. PMID:23593770

Satell, G. (2014). Five things managers should know about the big data economy. Forbes.

Say, J.B. (1803). Traité d 'économie politique: ou, simple exposition de la manière dont se forment, se distribuent et se consomment les richesses [Traduction on Political Economy: On the Production, Distribution and Consumption of Wealth]. New York: Kelley.

Say, J. B. (1815). De l'Angleterre et des Anglais. Paris: Arthur Bertrand.

Say, J. B. (1816). England and the English People (2nd ed.). London: Sherwood, Neely and Jones

Say, J. B. (1839). Petit volume contenant quelques aperçus des hommes et de la société, 3e édition entièrement refondue sur les manuscrits laissés par l'auteur et publiée par Horace Say. Paris: Chez Guillaumin, Libraire.

Schaupp, L. C., & Bélanger, F. (2014). The value of social Media for small businesses. Journal of Information Systems, 28(1), 187–207. doi:10.2308/isys-50674

Schumpeter, J. A. (1934). The theory of economic development: an inquiry into profits, capital, credit, interest, and the business cycle. New Brunswick, NJ: Transaction Publishers.

Schumpeter, J. A. (1954). History of Economic Analysis. New York, NY: Oxford University Press.

Sedkaoui, S. (2017). The Internet, Data Analytics and Big Data. In Internet Economics: Models, Mechanisms and Management (pp. 144-166). Bentham Science Publishers.

Sedkaoui, S. (2018a). Data analytics and big data. London: ISTE. doi:10.1002/9781119528043

Sedkaoui, S. (2018c). How data analytics is changing entrepreneurial opportunities? International Journal of Innovation Science, 10(2), 274–294. doi:10.1108/IJIS-09-2017-0092

Sedkaoui, S., & Monino, J. L. (2016). Big data, Open Data and Data Development. New York: ISTE-Wiley.

Shane, S. (1996). Explaining variation in rates of entrepreneurship in the United States: 1899-1988. Journal of Management, 22(5), 747–781.

Shane, S. A., & Venkataraman, S. (2000). The promise of entrepreneurship as a fi eld of research. Academy of Management Review, 25(1), 217–226.

Soyibo, A. (1996). Financial Linkage and Development in Sub-Saharan Africa: The Role of formal financial institutions in Nigeria. 001 Working paper 88. London: Oversees Development Institute.

Stevenson, H. H., & Jarillo, J. C. (1990). A paradigm of entrepreneurship: Entrepreneurial management. Strategic Management Journal, 11, 17–27.

Stokes, D., & Wilson, N. (2010). Small business management and entrepreneurship. EMEA: Cengage Learning.

Storey, D. J., & Johnson, S. (1987). Are small firms the answer to unemployment? Employment Institute.

Wiklund, J. (1999). The sustainability of the entrepreneurial orientation-performance relationship. Entrepreneurship Theory and Practice, 24(1), 39–50. doi:10.1177/104225879902400103

Zahra, S., & Covin, J. G. (1995). Contextual influences on the corporate entrepreneurship performance relationship: A longitudinal analysis. Journal of Business Venturing, 10(1), 43–58. doi:10.1016/0883-9026(94)00004-E

Zhu, Y., & McKelvey, M. (2013). Business models in Big Data in China: Opportunities through sequencing and bioinformatics. In How Entrepreneurs Do What They Do: Case Studies of Knowledge Intensive Entrepreneurship. Cheltenham, UK: Edward Elgar Publishers. doi:10.4337/9781781005507.00024

KEY TERMS AND DEFINITIONS

Business Model: A business model is a company's plan for how it will generate revenues and make a profit. It explains what products or services the business plans to manufacture and market, and how it plans to do so, including what expenses it will incur.

Customer Relationship Management (CRM): Is a business strategy that optimizes revenue and profitability while promoting customer satisfaction and loyalty. CRM technologies enable strategy, and 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.

Data-Driven Business Model (DDBM): Has become an ever-more important area of study and application and puts data at the center of value creation.

Entrepreneur: Entrepreneurship is not only an outcome of the ecosystem but also an important input factor since entrepreneurs drive the ecosystem by creating it and keeping it healthy. Drucker believes that what entrepreneurs have in common is not personality traits but a commitment to innovation. For innovation, the entrepreneur must have not only talent, ingenuity, and knowledge but he must also be hardworking, focused, and purposeful.

Entrepreneurial: A process in which opportunities for creating new goods and services are explored, evaluated, and exploited.

Entrepreneurial Activity: Entrepreneurial activity, as an output of the entrepreneurial ecosystem, is considered the process by which individuals create opportunities for innovation. This innovation will eventually lead to a new value in society and this is, therefore, the ultimate outcome of an entrepreneurial ecosystem while entrepreneurial activity is a more intermediary output of the system. This entrepreneurial activity has many manifestations, such as innovative start-ups, high-growth start-ups, and entrepreneurial employees.

Innovation: It is recognized as a source of growth and competitiveness. The Oslo Manual distinguishes between four types of innovation. Product Innovation: Introduction of a new product. This definition includes significant improvements to technical conditions, components or materials, embedded software, user-friendliness, or other functional characteristics. Process Innovation: Establishing a new production or distribution method, or significantly improving an existing one. This notion involves significant changes in techniques, material and/or software. Marketing Innovations: Establishing a new marketing method requiring substantial changes in a product's design, conditioning, placement, promotion, or pricing. Organizational Innovation: Establishing a new organizational process in practices, workplace organization, or company public relations.

Internet of Thing (IoT): The inter-networking 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.

Key Performance Indicator (KPI): Is a high-level measure of system output, traffic or other usages, simplified for gathering and review on a weekly, monthly or quarterly basis. Typical examples are bandwidth availability, transactions per second and calls per user. KPIs are often combined with cost measures (e.g., cost per transaction or cost per user) to build key system operating metrics.

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

Return on Investment (ROI): Is a performance measure, used to evaluate the efficiency of an investment or compare the efficiency of a number of different investments. ROI measures the amount of return on an investment, relative to the investment's cost. To calculate ROI, the benefit (or return) of an investment is divided by the cost of the investment. The result is expressed as a percentage or a ratio.

Small Business: Are companies that fall under specific legal limitations regarding the number of employees and the annual turnover. However, this differs from one country to another.

Startups: The first thing that is associated with entrepreneurship is startups. It is necessary to establish its definition in order for it to be used later on in this study. A startup is a human institution designed to create a new product or service under conditions of extreme uncertainty. A startup is also considered as an organization formed to search for a repeatable and scalable business model. The term scalable suggests that the aim of every startup is to grow (and, consequently, to stop being a startup) and to mature into a fully functional company: to an SME.

Time-to-Market: Is the length of time it takes from a product being conceived until its being available for sale. It refers to the amount of time it takes to design and manufacture a product before it is available to buy. It is important in industries where products are outmoded quickly. A common assumption is that Time-to-Market matters most for first-of-a-kind products, but actually the leader often has the luxury of time, while the clock is clearly running for the followers.

Chapter 8
Plan and Rules for Data Analysis Success: A Roadmap

ABSTRACT

Adapting the complex big data into your projects will be one of your strengths! Your mission to integrate big data is not limited to the use of sophisticated tools to solve your problems, but you must align the requirements of your activities with data lake or data warehouse through clear and correct strategies, taking into account your business as a goal. This provides support to your companies in all stages of your projects: from defining and taking requirements to start production and subsequent maintenance. Finally, it will help you create sustainable and stable competitive advantages.

INTRODUCTION

My freedom thus consists in moving about within the narrow frame that I have assigned myself for each one of my undertakings. . . . Whatever diminishes constraint diminishes strength. The more constraints one imposes, the more one frees one's self of the chains that shackle the spirit. Stravinsky (1942, p 65)

Today we are witnessing a strong enthusiasm around the theme of big data. Publications of different natures and demonstrations multiply and the promises also, without really defining the outline of the phenomenon, to be able to

DOI: 10.4018/978-1-5225-7609-9.ch008

approach it as a real project and not as a fuzzy and ephemeral technological mode.

Big data is an extraordinary opportunity for a company, a sector or even a country. Indeed, it allows having the useful and necessary knowledge at the right time to better manage the growing complexity of the operational. To take full advantage of this large amount of data, the first step is to define the process by which data should be collected, processed and analyzed. Then, we must identify the most appropriate business domain to launch a pilot or a Proof of Concept. Subsequently, we must validate the choice of tools and appropriate technologies and finally build an organization and governance to sustain and enhance the big data initiatives.

If some companies are already engaged in big data initiatives, the difficulty for others, especially small businesses and entrepreneurs, is how and where to start? Here are the important keys to implement when starting a big data project.

These different points that we put forward will allow future entrepreneurs to better understand the experience of value creation based on the big data analytics, as a whole. These tools will shed light on the conditions of success for entrepreneurship in the big data universe and on the different actions to be implemented.

DATA ANALYTICS WORKFLOW

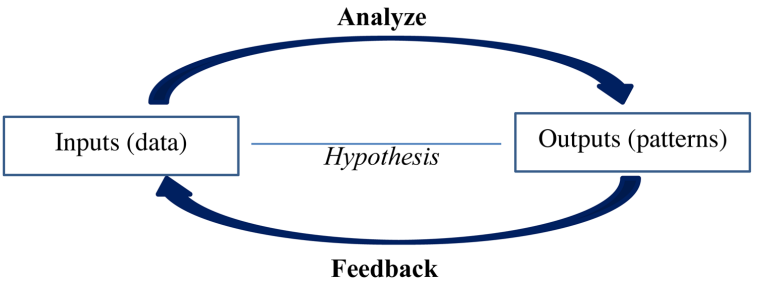
Data Analytics, big data, and machine learning are very popular terms in today's business world. However, perimeters encompassed by each of these terms overlap meaning different things. From a data point of view, big data refers to several Vs, in addition to the three famous Vs, which highlight the ability of traditional tools to process and analyze the available data (collection, storage, analysis, integration, etc.).

In this chapter, I will start by describing the workflow that can be adopted to better explore the data. Typically, I will explain how the data analytics process can be applied when working with big data in general.

Let's go!

But it should be noticed that, due to its experimental side, the data analytics will empirically run this workflow. As a result, this experimental model of work is not linear, but iterative: The analyst or the entrepreneur, who want

Figure 1. Data analytics process



work with data, will define a hypothesis, implement it, and then refine it. Usually, a big data analytics process is represented in the following form.

If you decide to work with data and launch your proper big data project, you need to have a clear idea of the implementation process to be performed because there are several steps to respect. From the setting up of good questions and the definition of goals to the exploration of the data through the preparation (collection, cleaning …) of that data, until the critical analysis of the results, globally, here is a data analytics workflow:

Definition and Formalization of Your Goal and the Business Need

Define the goals behind big data project means ask the right question for which we will explore answers. In fact, the data especially the unstructured data do not contain transparently, answers to everything. You have to know what you are looking for to reach your objectives. So, data analytics process is not a kind of "crystal balls" ... there is a great work behind the scene which needs to be done ... and the first step is to know what we are looking for!

For example, what do customers talk about on social networks about one of our specific products ... So it is not about an opinion poll ... but to spot just keywords in their conversations and that could express their feelings, their expectations, their choices, etc. This is to say, before starting a big data project, you must get closer to the business context of your project in order to define the main problem and identify the needs.

At the end of this small prospection step, by considering what is observed and the answer to be formalized, you will have globally a vision of what the job seeks to accomplish.

At this stage, you will translate the need into a more rigorous formalization. At first glance, you will have a more or less fine idea of the data likely to

solve the problem. Define rigorously a need to be solved by the data analytics process means also define what we are trying to predict: the target.

After defining the target, and taking into account the various business constraints, you will be able to decide the context of your project structure.

For example, if we want to develop a detection system of unsubscribing a subscriber to an online service, it will be necessary to define the notion of "the interest" of a subscriber to this online service. As such, you can formulate the following hypothesis:

- If the customer connects at increasingly spaced intervals, it means that his interest in the online service is weakening
- If the connection sessions to the service become shorter, we can deduce that it begins to lose interest in the online service.

In the light of these assumptions, you will be able to deduce the good 'hypotheses' in turn, in particular, on the data that will help you to model the phenomenon of disinterest. It may, for example, decide to look at historical data connection to the website (duration, location ...) and the activities of subscribers during these connections.

This initial phase focuses on understanding the objectives and requirements of the big data project from a business perspective, and then converting that knowledge into a definition of the data analytics problem. Here it is necessary to evaluate what are the expected gains and costs of the project.

Data Collection

After the rigorous formalism modeling of the business need, comes the phase of data collection. Collect the data i.e. search the unstructured data necessary for your business activity in conversational universes. This work is not always easy because most companies have never sought to collect unstructured information. Big data technologies are there to help ... but it also takes a lot of common sense and tips to know where to look for the data. Example: Facebook and Twitter are fields of exploration for algorithms called "bots", looking for keywords related to clients by browsing the web pages. The feedback will be occurrences of links, therefore, 'the meaning'.

You have to pay a great attention to this phase because it is a fundamental phase. It is long and expensive, but essential to have the guarantee of the reliability of the data before launching any analysis.

As data analytics is precisely focused on the data, the latter remains the nerve of war of this field. Depending on the business hypothesis that you would have made on the problem, you will look within for the data that will be relevant to support your hypothesis. At this level you have:

Internal data: Or the data you own, this can be databases, different digital files, emails, log files, photos, archives etc. You will look in these data sources to find the data that interests your project. Generally, the internal data reseal a lot of information, because they concern the internal context of your business.

External data: Coupled with the internal data, the external data can have a real added value. Data external concerns all data that is not produced by your business activity. Among these data, we can distinguish: social networks, videos on platforms like YouTube, tweets, Likes Facebook, etc. Moreover, these raw data contain a lot of noise and are generally more difficult to exploit.

This type of data can help you to better understand the interaction of your business with the outside world. In particular, you can use it to better measure the adherence and satisfaction of your audience and your clients. Such a measure is generally possible thanks to the NLP (Natural Language Processing), Sentiment Analysis and Machine Learning techniques.

Clean and Prepare the Data

Taking advantage of all this data can be a real challenge for you. Some data will be simple to use, such as those from relational databases. In this case, we speak of structured or semi-structured data. Even if we are talking about unstructured data, we still have to start preparing them. The data must undergo several "cleaning" before being really usable. Unstructured data, such as sounds, images, social signals (comments, tweets, etc.) are more complicated to process. In fact, this phase corresponds to transforming so-called unstructured data into quasi-structured data or totally into structured data! In particular, it will be necessary to apply some specific analysis to draw interesting indicators. Cleaning up the data includes several things, such as:

Remove Missing Value

Deleting missing value removes incomplete and inconsistent data in a dataset. Their removal will allow for a better performance of the predictive model. When we want to remove the missing value, we have two choices: (i) deleting them simply or (ii) replacing them by default values. Simply removing missing data can occur when the absence of certain values does not allow

for a tangible observation. In this case, we delete all the observation because it does not bring a real added value.

In some cases, it is quite possible to catch up by repairing an observation with missing data. This is possible by replacing them with default values. For example, for a dataset of observations on the Germans people, we can have several characteristics (features). Notably: gender, size, weight, occupation, salary, etc. If for an observation, it misses the 'size' feature, we can assign a default value that will be the average size of Germans for example.

This type of data processing requires an understanding of the business to assign logical and adequate default values to the phenomenon to be modeled. This will avoid introducing a bias that will distort our model.

Remove Outliers

An Outlier is a disproportionate value that comes out of a certain range for dataset observations. For example, if we have a dataset about the salary of several people in Germany. Suppose that in this dataset, we observe an average salary of 2000 € per month and that the variation of these is between 1500 € and 4000 € monthly. If we observe for one person that he has a salary of 30000 € monthly (a football celebrity for example), we can say that it is an outlier.

In addition, outliers also concern abnormally low values. In the end, an outlier is any value that is abnormally different from the set of observations. Removing these outliers, before applying the machine learning algorithms or any analytical technique is essential. Indeed, it allows not to biasing the predictive model obtained. Indeed, the goal is to obtain a statistical model that generalizes well to the real situation. And not one that adapts to outliers.

Data Exploration and Interpretation

Even before being able to process data, it will be necessary to put them in the file and store them. This phase obviously makes the data already much more presentable, since they will be at the same time formatted and labeled (indexed). Then we can start the exploration phase, which can be considered as the actual processing phase that will give the sense to the data. The data rarely give correlations or conclusions. It is necessary to know how to interpret them, to do so you need to have a critical view.

The data exploration phase provides an understanding of the different techniques and methods of analysis. Understanding the data is like

understanding the composition, the distribution and their interactions. One of the simplest methods for exploring data is to use the tools of descriptive statistics. In particular, indicators such as Mean, Median, Variance, and Standard Deviation, etc. These indicators give a concise picture of the distribution of a feature.

Also during the univariate study of features, visualization of data using tools such as histograms, pie charts, and box plot … is necessary.

Crossing the features with each other during visualizations allows you to glimpse less obvious relationships to be observed at first glance. 2D diagrams in the form of point clouds or even 3D diagrams make it possible to see the distribution of data in a multi-dimensional space. Finally, the purpose of visualization and data mining is to understand the data. Another purpose of these methods is to validate that our dataset is well cleaned and ready to be used for analysis.

Modeling

After the cleaning, data preparation exploration phase, comes the modeling phase. The goal of this step is to build a statistical model able to 'predict' the result of a given phenomenon (problem). This model will be based on a dataset representative of the phenomenon we are trying to model. Machine Learning, for example, will learn from these data to build a statistical model. This model will be used to predict the result of an observation you have not yet seen.

The goal of this phase is to build a model that is a good approximation to describe the real phenomenon we are trying to model. For this reason, you must try several hypotheses and test them to produce the best possible model. In this context you have to:

Test the Performance of the Model

After training and obtaining a statistical model, the question of performance of your obtained model arises. Indeed, we want to know how much the predictive model is generalizing on data that you have not seen. For that, you will use a Testing Set in order to test the performances of the predictive model. Performance calculation should quantify how well your model behaves. The idea is that this metric is a concise and easily interpretable value for you to know how the model reacts. Also, do not hesitate to

present the problem differently and to test other analysis techniques before committing to a result.

Optimization of the Model

Finding a coherent predictive model that becomes generalized well is an iterative and empirical process. For this, you will go back and forth between the modeling phase of the predictive system and that phase of measuring the performance of your model. During these iterations, you will further refine your hypothesis about the data, as well as the features that come into play in the prediction. During this phase, you may need to search for more data if you realize that you do not have enough data. In this case, data preparation and exploration phases are necessary to integrate the new data obtained in the model.

Deployment Phase

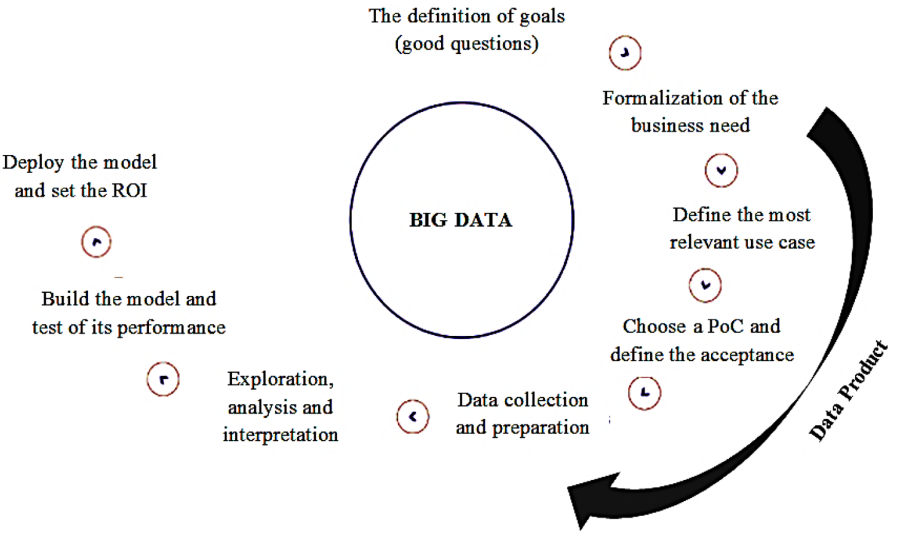
After obtaining a satisfactory model, comes its deployment phase as the end of data analytics workflow. In previous phases, generally, you work on relatively small datasets. In order to quickly test your ideas, you will use, for example, scripting programs like Python or R. These programs show their limitations when dealing with large amounts of data. This concerns, in particular, the production environments. In this case, a rewrite of the obtained model in a more robust software infrastructure is required.

Generally, to analyze the big data many architectures can be used. Predictive models are written using technologies, such as Hadoop, MapReduce, Spark or Mahout … This work of rewriting can be entrusted to data engineers who will be better placed than you to master this kind of technology. After the rewrite and production deployment phase on big data infrastructures, you will be able to observe the system performance in a real situation. This may give you, possibly, new ways to improve.

I do not know if you have realized, but we have just reviewed the workflow of a real data scientist's work and how he is doing to build a data product based on data analysis techniques coupled with machine learning algorithms. This workflow is empirical and the improvement of a data product is iteration after iteration.

A data product is an asset that relies on data and processes it to generate results using an algorithm. The classical example of a data product is a

Figure 2. Data analytics roadmap



recommendation engine, which ingests data and generates personalized recommendations based on that data.

Among the most relevant concrete examples, we can cite Amazon's recommendation system or Netflix's recommendation system. Similarly, Gmail's spam filter is a data product, since an algorithm is responsible for processing and analyzing incoming emails and determine if it is a spam or not. Computer vision, used by autonomous cars, is also a data product. Integrating machine learning algorithms are able to recognize traffic lights, detect other cars or pedestrians etc.

Bringing the big data to business requires a comprehensive systems approach, inclusive of intelligent processing and sensing technology, connectivity, software, and services, along with an ecosystem to address the smart environments of big data applications. The ultimate goal of any ecosystem is to evolve from being subsidized until it becomes market funded and financially self-sustaining (Sedkaoui, 2018c). Silicon Valley is at this maturity level Harrington (2017).

This development roadmap (see Figure 2) indicates that the progress in relevant technology will continuously contribute to the development of big data application. Therefore, we argue that big data project is promoted by business driving forces of both technology push and market pull.

Moreover, the type of analysis which is needed to be done on the data depends highly on the results to be obtained through decision-making. This can be done to (Sedkaoui, 2018a):

1. Incorporate massive data volumes in analysis or;
2. Determine upfront which big data is relevant.

So, we have two technical entities 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 (Shroff, 2013; Siegel, 2016).

KEY FOR SUCCESSFUL BIG DATA PROJECT

To become operational in the big data context, and in order to better guide your project, several keys are imposed. We will discuss here the good practices that can lead you to a successful big data project, whatever the size of your company.

It Is Not About Big Data, But Everything Is About the Value Generated From It

The situation, nowadays, is that most projects are not big data, in the sense that the data is not necessarily big. On this aspect, I am even unable to define what is meant by 'big' which is often referred to as the volume of the available data. It is from how much (GB, TB, or ZB) we can say that we are talking about big data? Moreover, for example, your small business activity, which is limited in size, budget, IT tools … compared to American, German, French, Indian, Chinese companies, can it in big data situation?

So, what is important is not the size or technique, but rather how to generate value. Analyzes, correlations test, predictions, machine learning algorithms ... The generalization of big data analytics in daily business topics, refining its work thanks to a contribution of statistics, in short, being data-driven, this is where the value of big data and all its interest lay. This may surprise some, it will reassure others, and we could do this for years. Today, data, especially external data which take unstructured formats, are just more available and,

both economically and technically, the analysis is more accessible. Understand that big data opportunities do not lie in the volume of data, but in the digital transformation of your business processes, is very important.

The big data revolution is explained by two phenomena. On the one hand, the nature of the data has changed. The source is not only the companies but also the users (customers), who now generate content: texts, photos, videos... and even the machines (sensors …). So making available a volume of data and a variety completely unpublished. On the other hand, the dynamic exchange has exploded the speed at which these contents are broadcast, since everything is now shared on networks: Facebook, Twitter, Instagram, etc. These two phenomena explain the famous 3Vs of big data.

But globally, big data is seen by everyone as the explosion of data volume, the phenomenal size of the data produced by the advanced IT tools. For the more technical professions, the big data refers to the several Vs, already discussed in section one of this book. Moreover, this perception of big data is so steeped in manners that the experts have decided that we associated the 3Vs when we defined this phenomenon. Unfortunately, designing big data in a pure volume aspect can:

- Minimize the potential of the data for business activities;
- Limit your perception of the digital transformation that is going on;
- And miss the reality about what big data really hides;

If you want to succeed in your big data projects, you must understand that big data is not only about volume, it is a social phenomenon. This is the visible part of the world's transition from the industrial age to the digital age. More concretely, it is the expression of two factors: (i) the provision of the internet and (ii) the increase in the number of people connected. Indeed, the popularization of the internet has led to the digitization of business activities, which has occurred at the same time as the increase in the number of connected people, through divers' sensors. For example, today, people are using more and more their smartphone (surf, purchase, social networks …), they are also conducting smart cars, and they live in the smart environment etc. All these digital activities generate data.

If you combine all these activities with the number of connected objects (IoT), you will find yourself facing an unprecedented amount of data. But this is not the point that I want to show, because big data is just a result, a consequence and not a cause. Thus, if you plunge in your big data project taking into account only the volume aspect, so you have to know that you

are attacking the consequence and not the cause. So, if you want to succeed in your project in big data context, change your angle of attack! Be more interested in how you will monetize this volume to better take advantage and create your proper business value.

Defining the Goal

We can never say it enough: "*there is no good wind for those who don't know where they are going*". The success of your big data project is associated with the level of clarity of what you want to achieve with this project. This is not just valid in a big data context but in any context. If you want to undertake in the data revolution you must define the goal behind your big data project.

Big data opens up a vast field of possibilities for entrepreneurs, but at first, it is important to define the use case on which an entrepreneur must focus its energy and resources. For example:

- **Create new Services:** Capitalize on connected objects to bring new services to consumers and develop tomorrow's business models.
- **Improve the Customer Experience:** Improve the customer's understanding, preferences, behavior, through the analysis of their transactions, multichannel interactions and actions on social networks. You then can have a 360° view of your customer, this global view helps you to create a personalized marketing campaign.
- **Reduce Attrition and Build Customer Loyalty:** Reduce the cost of customer acquisition and retain existing customers through real-time event analysis. Thus, the entrepreneur can define the best next action to engage with customers and improve their satisfaction and maintain them.
- **Improve Operational Efficiency:** Transform financial processes, improve performance management, through the analysis of a large number of data and the setting up of dashboards to ultimately, take faster the best possible decisions.
- **Improve Security:** Reduce risks, fight against fraud, monitor in real-time through the examination of large amounts of internal and external data to identify anomalies.

The definition of the goal that you want reach by embracing big data universe means that you should not do it just to prove that you too can do it. You just have to be aware that big data is the expression of a potential

phenomenon and that the real challenge of a big data must focus on how to transform your business to generate value.

The challenge for companies today is to go beyond the wishful thinking of big data and to succeed in implanting it sustainably. You have to know that it is a project like any other; it requires setting the goal, or defining the financial, human and logistical resources … that will be needed to achieve this goal. This is to say define the scope of impact and if you are not yet aware to define the goal behind your project you can start with a PoC (Proof of Concept) and using an iterative method, refine your goal as you go along with the results and directions you get from this PoC.

In fact, paradoxically, most organizations do not yet benefit from their investments; many are struggling to finalize their PoCs as the proof of their feasibility. In any case, it is very rare to have a very clear vision from the beginning; generally, the vision becomes clear as one progress.

Preparation and Planning

Big data can bring huge benefits to businesses. However, as with any business project, preparation and good planning are essential, especially with respect to infrastructure. Until recently, it was difficult for companies to engage in big data without making heavy infrastructure investments (expensive data warehouses, software, specialized analysis staff ...). But times have changed. Cloud computing offers many options for big data and this means that businesses can tap into their data without having to invest in huge storage and data processing instances.

To advance in the big data universe and turn data into knowledge and generate value for your business, you will need to pay attention to the following key infrastructure elements:

Data Collection

Whether produced or collected by your company, data comes at a specific point in your information system. Customer database, consumer comments on the e-commerce site and on social media, marketing lists, and emails archives; the data is increasingly massive and heterogeneous. And often you have to buy or retrieve external data to enrich your internal analysis.

If you need to get new data, your need for new infrastructure may increase dramatically. But all this naturally depends on the type of data you need. The

data can come from sensors positioned in the devices, machines, vehicles etc. From applications, for example, it may be an ordering application for products intended for customers. The data can also come from other sources, such as surveillance camera circuits for example.

With a little technical knowledge, you can set up these systems by yourself, but you can also use the service providers to set up these systems and collect the data for your business project. Collecting external data sources, for example from social media, does not require in-depth changes to your infrastructure, since you have access to data that someone else collects and manages. If you have a machine and internet connection, you are pretty much equipped to start.

Data Storage

As the volume of data generated and stored by companies has exploded, sophisticated yet accessible systems and tools have been developed to assist you in this task. The main storage options are a traditional data warehouse, a data lake, a distributed or cloud storage system, and of course your enterprise server or the hard drive of your machine.

The classic hard drives are now available with very high capacities and at low cost. If you conduct a small business, it may be all you need. But when you start processing a large amount of data for storage and analysis, or if that data is destined to become a key part of your strategy, you need something else. A distributed system like Hadoop, or cloud-based, might be better suited.

In fact, a cloud is an attractive option for most businesses. It is flexible, you do not need physical systems, and it reduces your need for security tools to protect your data. It is also a lot cheaper than investing in dedicated systems and data warehouses.

Data Analysis

If you want to use the data you have stored to find something useful, you will need processing and analysis capabilities. So, all the challenge of this brick is to transform data into knowledge. This is where the programming languages and the analysis platforms join the game. Three phases, which were more detailed previously, for completing this process:

- Data preparation (identification, cleaning … to make data ready for analysis).

- Construction of the analytical model.
- Draw conclusions from the acquired knowledge.

IBM, Oracle or Google, and many other companies offer tools to turn raw data into information. 'BigQuery' of Google is designed for example to allow anyone with a little knowledge about data to run large data sets. 'Cloudera' and 'HDInsight' are also in this niche. And many startups also offer solutions to this end.

Data Security

To optimize your performance, you must be aware of the various threats that affect your data. Developing an accurate view of data and hierarchy allows you to better secure critical information for your business, and provides some agility to better adapt the myriad of rules and regulations that you need to comply with.

Data Visualization

Too often companies are burying real nuggets of information that if brought to light could really affect their strategy. Why? Because this information is embedded in a fifty (50) pages (report) or a graph so complex that no one can understand it. If key information is not clearly presented, it will never be used to initiate action plans. Data visualization is necessary to better understand the trends.

The main types of data visualization include dashboards, commercial data visualization platforms, but also simple charts and tables that allow you to communicate ideas quickly. Most small businesses looking to improve their decision-making can rely on simple graphics and visualization tools like word clouds or even Excel.

Resist to the Changes

One of the biggest stumbling blocks to the success of your big data project or the success of any other project is the resistance to the change. We are naturally averse to change; this is often due to the weight of habits. Big data represents a change, the change from one era to another, the transition from an industrial era to a digital era. This change is inevitable and will be done with or without our agreement. If you want to successfully guide your big

data project you have to avoid what we call the frog behavior. It refers to a type of resistance to change that denies reality. Kodak and Nokia are two examples that have this behaved.

Kodak wanted to impose its vision of the image over time and denied the reality of the digital image worn at the time by companies like Xerox and Fujifilm. Unfortunately, this relentlessness to continue working with a product from another era cost him his life. Nokia, also, underestimated seeing the progressive impact of the iPhone launched by Apple on culture and thus missed the train of smartphones. Like Kodak, this error was fatal and in November 2011, the company was bought by Microsoft.

Big data is the expression of a change that will force all economic actors to review how they create value and boost their business model. Launch a big data project means undertaking an approach that engages the activities, and the company, in an area where the culture of the company is driven by the data.

Old BI approaches tend to focus stakeholders on the tree, which sometimes obscures the forest. The big data approach is much more interested in the forest itself. It seeks to offer a business landscape by giving perspective. It involves both intensive sharing of information and decision processes guided by the data, more precisely by the patterns, knowledge and predictive models derived from the data analysis.

Predictive modelling and analytics can be of crucial importance for entrepreneurs if correctly aligned with their business process and needs and can also lead to significant improvement of their performance and quality of the decisions they make, thus increasing their business value (like: Amazon, eBay, Google, Facebook …) (Davenport and Kim, 2013; Siegel, 2013).

Success is subjective; to be fulfilled in professional life and build your own big data project, it is essential to be able to ask the right questions, yes! But also it is important to build a good plan considering the following elements:

- Success is driven primarily by the quality of the business model. It is the combination of an agile technology foundation, advanced analytical skills, robust data management procedures, and most importantly, centralized governance and the durability of the activity.
- Big data is an excellent field of experimentation for new working methods. As soon as business use cases have been prioritized, the first challenge must be to test the approach with feasibility demonstrations (PoCs), to demonstrate the potential of a big data initiative, or to negate it by learning collectively about its failures.

- You must succeed in aligning your business project on the potential of big data as a lever for innovation and value creation, both for business and IT, in order to secure investments and co-deliver the roadmap.
- Astringent data security policy is required. This last point is a key success factor from the point of view of companies.

It is thanks to these different pieces of the puzzle that big data will be able to keep its promises. Also, from experience, I have found that a data-driven project that does not have a sponsor is very unlikely to be deployed. Hence, the importance of having a sponsor can embody your big data project. As in all transformation projects, you must show results quickly. Achieving a Proof of Concept or a "quick win" approach is a good way to prove the value of your big data initiative. These results will help you to overcome reluctance and resistance to change and ensure the progress of your project.

GOOD PRACTICES TO ENSURE THE SUCCESS OF A BIG DATA PROJECT

Harnessing huge amounts of data can dramatically improve the performance of your business, but you will have to review your decision-making process. This statement, which is credited to E. Deming as well as to P. Drucker, the two management gurus, explains why the recent explosion of digital data is so important.

Data is everywhere, it is a reality. But how should companies collect it, and analyze it to generate value? The various steps detailed throughout this book proposed a grid of several axes of analysis and can help you to understand the different applications of big data. However, this grid, even if it allows you to structure the applications of big data (**why?**), but it cannot clearly guide you to find the way: (**how?**) that will allow you to put these applications in action. It is, therefore, necessary to detail the way in which big data are implemented in practice.

To put it simply, big data offers businesses and decision-makers significantly increased means to measure, and thus better know the driving forces of their activity, which allows them to improve both their decision-making process and the performance of their activity.

Let's take an example to better clarify. Historically, the booksellers were able to distinguish which books were selling well and which is evil flowed. By

setting up a customer loyalty program, they could associate certain purchases with certain buyers. But, as soon as the purchases were made online, the profile of the customers has been considerably refined.

Not only that they could track the purchases, but they also knew what had interested the customers, what journeys they had made behind the screen, and how promotions, comments, and page layouts had influenced them. They also found similarities between some customers or groups of customers. Soon, they have developed algorithms to predict which books their customers would like to read, these algorithms refined each time the consumer followed a recommendation or ignored it. For their part, traditional booksellers still have no way to access this information, or how to use it.

As we have seen in detail, throughout this book, the scope of the big data revolution is incommensurate with traditional data analysis. The traditional approach of the data management is to centralize the storage and processing of data in a central server in client/server architecture. This data is managed in the server by a relational databases management system. Unfortunately, traditional data management approaches are finding it increasingly difficult to adapt to the big data constraints that are new.

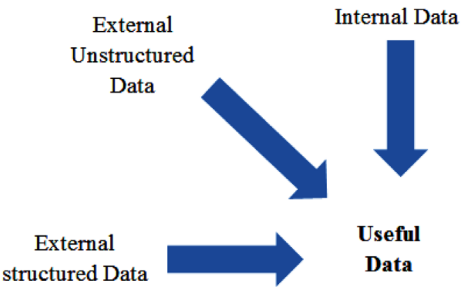
To manage the challenge behind this data revolution, the technological approach is no longer to centralize the storage and processing of data in a single server but to distribute the storage and parallelize the processing of this data on several machines. Hadoop is the most mature software implementation of this technology approach. It is used today at least as a pilot by all companies wishing to engage in the large-scale to exploit their data.

With big data technology, it is now possible to measure more precisely and therefore to manage more precisely. In this way, we can make better forecasts, make more relevant decisions and better target our initiatives, including in areas where intuition have so far prevailed over information and rigor. As the tools and philosophy of big data spread, the pre-existing ideas about the value of the experience, the nature of the expertise and the management methods will change.

But, the question that arises, especially for entrepreneurs or small businesses is: How to exploit big data?

The answer to this question lays in the fact of well detailed the analytics process of these big data. This requires:

Figure 3. Product useful data



Listing the Usable Data

Internal data, external data, structured and unstructured data, the nature and sources of the data that will use to achieve your goal are crucial. This has a direct impact on the technologies to be put in place and on the quality of the results. The first source of data that can be used by you for your business project obviously covers those that have been generated and stored we are talking about the internal data, such as: for example the data about purchases that can be used to analyze customer behavior (via CRM) and other data on internal processes from ERP management tools.

Then you have also the external data, which is not generated by your business activities, but which is available indirectly usable formats: open data, survey data or panel, etc.

Also, do not forget to integrate unstructured data scattered throughout the digital world and which accounts for more than 85% of big data. For example, data from social networks, recommendations, opinions, etc.

Unlike the data generated by your activities (internal IT systems), these data are unstructured because they are not directly readable according to a predefined structure. It will, therefore, be necessary to clean them so that they can be analyzed. Once your data are outsourced and ready to be used, you have to know how to value them.

If the goal is to explore and analyze social networks data, simple-to-use solutions like Watson Analytics for Social Media may be enough. On the other hand, if it is necessary to correlate information of the web with internal data, to include open data, or else flows of information coming from connected objects, a deeper reflection on the implementation big data technology is necessary. This often involves several responsibilities: how to manage data collection, lifecycle, security, and access ... how open

source and SaaS solutions can help create an open and less expensive platform, etc.

Measure the Evolution of Your Activity and Understand Its Causes

Most modern companies are able to measure the performance of their various activities. These measures can lead to reports to control and manage the marketing, operations, finance, or accounting. This is the basis of quantitative analysis. The most famous example is probably the Walmart sighting, known since the early 90's; that between 17h and 19h, customers have a propensity to buy beer and diapers for children. Walmart has reorganized its offer to better serve customers.

Then, comes the last level of maturity in the data analysis process which concerns the understanding of the relationship between the causes and effects of an activity, and to be able to draw conclusions about its trends.

It is this inductive and predictive part of big data analysis that has given it the power and makes it in the heart of advanced analytics. Indeed, it allows on the basis of factual observations to draw conclusions about the general dynamics of an activity and thus to establish scenarios, to anticipate the events based on the model, and thus to optimize a business.

The Choice of Technology or the Toolbox for Data Analysis

Data analysis is based on technologies. Without them, nothing is possible. Any data analysis project starts with the choice of technological tools. I will group them into three categories, corresponding to three types of needs and levels:

- **Tools for Beginners:** You are just starting to analyze data for your entrepreneurial activity, your start-up or your e-commerce site for example. To dive into the world of data analysis, you can use Google Analytics, Google Tag Manager (GTM), Regex101, Excel …
- **Standard Tools:** You have more budgets for your big data project and you would like to do a more detailed analysis of your data. These are advanced level tools that are used by data analysts. They can meet more specific needs than the previous tools, like Optimizely, Dataiku DSS, Crazyegg, Mixpanel, etc.

- **Technology for Experts:** You have a team dedicated to data analysis and you want to exploit your data with very specific needs. From a certain stage of advancement in data analysis, it is preferable to abandon the software "easy-to-use" analysis or testing to create your own tools (scripts, pipelines, automation ...). It involves immersing yourself in the code. There is a beginning to everything, but do not care because it is less complicated than it looks. Among these reference tools in data analysis, we can mention Hadoop, MapReduce, Spark, SQL, Python, etc.

If you are embarking on data analysis, we advise you to start by familiarizing yourself with the first family of tools (beginner level). It is useless to want to go too fast. The transition from 'beginner' stage to 'standard' or 'expert' (advanced tools) will be done gradually, as your needs become more specific or your databases grow. The most important thing is to know what you want to do with your data because that ultimately determines the choice of the technology.

Identify Solutions Through Data Analysis

It is about finding solutions that are reasonable for your entrepreneurial project, even if it means adapting other solutions to the resources and requirements of your business. It will first rework the data (preparation, cleaning …) so that they are usable for the intended application. The first step involves extracting relevant information from databases. In the case of a data warehouse, it will be necessary to fetch several tables and manipulate them in order to keep only the relevant information. In the case of unstructured data, for example, text or images, videos… it will be necessary to extract the attributes, like image size, subject, style … that are relevant to your business.

For example, if we are interested in the users' opinions, we can identify whether the opinions are positive or negative via text mining tools that can extract the negative or positive character of a text. Then, we have to select the most relevant variables and proceed to the assignments specific to the trade, calculation of financial ratio for example. In particular, in the case of the search for causes and effects, it will be necessary to identify the variables of effect, or dependent variable, from the independent variables (the cause), and those which will serve as variables of control (linear regression). This step is in practice the one that requires the most time.

Then, quantitative methods can be applied according to your project objective: from simple statistics to predictions via the recognition of structures.

This is the part that requires a good knowledge of statistical tools and machine learning algorithms (see chapter six). But, if the data has been well prepared, this step does not require much time to derive useful knowledge.

Once the amount volume of data has been prepared and analyzed, it will be possible to draw the business solutions: produce reports, identify new opportunities, and detect ways to make your business more efficient …

In this context, I want to draw your attention to an important point; it is necessary and important to go, in your analytics process, step by step. This is to say, that you have to do so by respecting your maturity regarding the data analysis.

In practice, it is always interesting to produce simple statistics when one is interested in a new problem, and, even if you cannot yet have the ambition to put in place predictive techniques, you can already be interested in some "basic" analyzes. There are also other conceptual frameworks that you can develop, and which we are detailed in this book.

In addition, care must be taken to follow these steps: for example, assigning missing values to a database without a good knowledge of it can cause a great risk and that may bias the results. It is better to take the time to move forward on a solid foundation and established skills, rather than drawing quick conclusions that do not rely on real analysis.

HOW TO GET STARTED IN DATA ANALYTICS?

When I meet my marketers' friend or others, they have a little more trouble perceiving what big data can change in their daily work. With which concrete case can I get started to differentiate my project? When attending meetups or other events around the data, many curious people ask frequently: how can I start in big data and join the data analytics arena?

The answer is simple it is about practice. Data analysis consists of studying and analyzing a set of data in order to transform them into added value for the business strategy.

On a typical project, we usually start by exploring raw data. An intuition or orientation (common sense, business purpose, etc.) will be needed to choose a direction. Then, it will be about cleaning and consolidating the data, building variables and establishing statistical models.

The entrepreneurs who want to work with big data must be very autonomous in his work, with his various skills (basic are necessary) in computer science, statistics and some knowledge about its business sector.

Of course, it is not the same for a data scientist or an expert in this field who must have very advanced knowledge in several areas including programming, data mining, machine learning, text mining, big data … but it is a bit complex for beginners.

In order to simplify the context, I have gathered in the following some key points which can help you to make your first steps in sweets in data analytics independently. It is more about resources for all interested people to better understand what lies behind the "hype" of big data, and also take pleasure in discovering this universe rather fun!

Statistics and Mathematical Basics Are Needed

Data analysis requires a certain affinity with statistics and mathematics. In any case, do not hesitate to review the basics of these two areas. Descriptive statistics, for example, are absolutely necessary and for everyone who tends to plunge in big data ocean. Mean, variance, standard deviation, confidence intervals, sampling methods, and analysis of variance … are part of the everyday data analysis process.

Also, Graphic representations (histogram, box plot …) are very important, especially to visualize your data in any stage of your analysis, this can help you too to understand your results and then derive insights.

Explore the Data: First Analyzes

In data analysis, there is not a technique or methods but several ways to proceed. It is sometimes said that it is "hack". So do not be afraid to explore or analyze our data following your own direction of work or do it your way.

To get started, why not begin with open and explore your data in a spreadsheet, for example, Microsoft Excel or Google Spreadsheet. Create tables to summarize data (counting values, data distribution, cross data and analyze the relationships in order to derive information) and graphs to visualize them are basic actions in any analysis process. Data analytics process requires a good overview of all data.

Spreadsheets, and in particular Excel, make it possible to carry out basic statistical operations. Take the linear regression as an example. I assume you already have some basic knowledge about linear regression, obviously after reading chapter six of this book, where concepts about several methods have been well detailed.

Graphically, the simple linear regression aims to find the linear equation that best presents the phenomenon studied. This very simple model already makes it possible to get good statistical approximations. We have also the multiple regression which explains a dependent variable according to several independents variables. It is also possible in Excel.

The spreadsheet files (csv, xls ...) are practical but only for small datasets and not requiring too much manipulation. Database work eliminates many limitations of Excel. We do not usually work on a workstation but on a dedicated server (or even a set) and this opens the possibility of processing large volumes of data, simultaneously and reproducibly.

In this case, there is a standard language for exploiting databases, such as SQL which makes it possible to query the database in order to find data, to cross them and to create insights. SQL is widely used in all fields, not only in computer science. Biologists, doctors, econometricians, marketers and all the professions that need to work on statistical data are forming there. It offers the advantage of handling large volumes of data, aggregations or calculations on the data without taking care of the technical implementation. There is no need to encode or find a powerful algorithm because it is not difficult for users.

Knowledge of Machine Learning Algorithms

After the exploration and preparation comes the modeling phase. The real value added on your data is obtained basically in this phase. To begin, it is good to have notions of statistical models and machine learning algorithms. For the perfect knowledge of techniques and mathematical formulas, we will leave this to a data scientist expert. But, for a beginner, or the entrepreneur who seek to learn from his datasets and generate value in order to boost its entrepreneurial activities, here are three methods that we commonly encounter and which were detailed previously:

- Regression such as simple linear regression, multiple linear regression …
- Classification: in order to classify elements into classes.
- Clustering: aims to divide a set of elements into homogeneous groups.

Let’s briefly illustrate what an algorithm can do with a simple case, probably closer to our everyday life: “an antispam filter”. At first, one can imagine that your machine will analyze how you will classify your incoming

emails in spam or not. Thanks to this learning period, it will deduce some big criteria of classification. For example, the probability that the algorithm classifies a spam will increase if the email contains terms such as 'money', 'offer' or 'win' and if the contact is not in your list. On the other hand, the probability of ranking in spam will drop if the contact is known and the words of the mail are more traditional.

If You Can See It, You Can Understand It: Data Visualization

It is well known that information obtained visually is easier to interpret, understand and analyze. The data visualization intervenes throughout a project on the data to support understanding. Building graphics or other visualizations helps to understand your data. Above all, data visualization allows the final phases to render, explain and highlight your work. This participates in the retransmission of the information. The easily accessible tools are rather numerous. In addition to software solutions (just Excel and Tables), many online sites allow creations in minutes. One example among many others: "RAW".

On the other hand, creating animated visualizations is more complicated. It will probably be necessary to use JavaScript and the "D3.js" library.

Think About Security and Privacy

Any big data project requires looking again at collected data. At a time when perimeter boundaries no longer have the same role and relevance, it is essential to protect the most sensitive data first rather than the access routes. In addition to the issues of authentication and access policies that need to be redesigned, big data also often require to take into account the compliance of data. Because the bulk of big data projects are consumer-centric, it is essential to adopt measures and behaviors that respect the privacy aspect. Opting for cloud solutions solves some problem.

Do Not Neglect the Real-Time Analysis

It is true that the aspects mentioned before are of paramount importance, but the main reason for the introduction of big data is the reliable and the real-time data analysis. Therefore, real-time analysis is an important aspect to consider when choosing a big data technology in your project. Also, a data analysis as fast as reliable could lead to a considerable reduction of the

costs to the delight of the customers and the company itself which will thus be highly competitive.

Python, Hadoop, Spark and More

In the past, to use data it was important to have deep knowledge about tools and methods, something that was not obvious to everyone. But with technological evolution, big data technology are more and more simplified to be easily used by people without deep technological knowledge. These tools eventually allowed everyone to see and question the data collected.

There are some essential tools today in the world of big data. This concern, in particular, the mastery of the R language specialized in the statistical analysis and/or the Python language and its associated libraries (Pandas ...). The learning curve of a programming language is a little harder but to better work with data you need also, as far as possible, to be a good programmer. This will make it possible for you to overcome all limitations in the collection and preparation of data, in the creation of variables and especially to generate and compare many models.

To work on large volumes of data, some technologies like Hadoop, MapReduce, and Spark … have become a standard. It is essential to master the framework in a professional setting. Also, with Hive, it is possible to manipulate data on Hadoop via SQL queries.

How to Start Immediately?

I hope these few tracks open your eyes to the different areas that are open to you. To get started, nothing better than training on data. This is precisely what I am talking about throughout this book, obviously by practices on concrete cases. All you need Excel or other big data tools (Hadoop…) to start! These tools can lower the barrier to entry for beginners, for example, manipulate different phase of the data analytics process, like data preparation and cleaning, data presentation in a visual interface, first predictive model easily achievable, etc.

In order to simply start and obtain the first success, it is better to identify a specific case and deliberately restricted, for example with CRM historical data that we control. This is to ensure a smooth start, results in the short

term, and a rise in competence. Finally, do not forget the fun side, or the one related to the practices cases out of business.

Big data is not just a vocation for the future. This is already a sector full of promise with many opportunities for growth. Many companies have realized the benefits of data analytics. If you are interested in such a field, it will first of all be necessary to arm yourself with courage. Because, the race to acquire the right competencies in this area doesn't seem to slow down while the labor market is unable to cope with an exponentially increasing demand (De Mauro et al., 2016).

Big data is a new sector, and no pre-determined path is being traced at the moment. There are, however, several ways and approaches to increase your chances of success for your big data project.

Basic Information

All good careers start with a solid foundation, namely a basic knowledge related to the big data and the algorithm applications. Unfortunately, in the case of big data, university courses are not yet numerous. Some institutions nevertheless offer their own degrees and programs aimed at training the next generation of big data experts. Learn about big data training program and choose the one that suits you best to acquire a robust foundation.

Big Data Technology

In many professions, work depends heavily on its tools or technology. In the case of big data, there are many technologies, namely software, that are essential to master. The technologies for setting up a data analytics process are multiple, you can implement the algorithms, detailed in chapter six with a programming language of your choice (JAVA, C ++ or others) as you can use libraries. Usually, the data scientist community uses one of the two programming languages:

- R which refers to a statistical environment easily allows the manipulation of mathematical functions and the graphical representation of the results.
- Python for its simplicity and the availability of libraries that implement the most used algorithms.

Thus, Apache Hadoop needs to be studied at length before being fully mastered. Its various components Hive, Pig, MapReduce, HBase, Spark, and many more must be understood in depth.

Skills

In addition to the technology, it is essential to gain some skills to succeed in your big data project. Techniques and algorithms like machine learning, data mining, and text mining … are essential for all the jobs which lie around big data. The same goes for data visualization, essential for communicating your discoveries or results generated by the data analytics process. In order to boost your big data project, you need also a creative spirit in your problem-solving path, because you will use the amounts available data to solve problems (the goal identifying in the first) and this is a faculty and always an asset and more.

Understanding of the Business Activity

Different sectors do not have the same uses or applications of big data solutions. A bank, for example, uses big data differently comparing to the medical institution. The same goes for manufacturing, security, transportation, education, or even sports companies. It is important then to know how several sectors can use these analytical techniques. If you have a preference, familiarize yourself with the sector in question and develop expertise in addition to your big data skills.

Be aware that large volumes of data are not a problem for your work and your business activities, but instead a great opportunity to find valuable information. You have to know that behind the data, we have correlations ... more or less certain!

Whatever the algorithm you will adopt for the analysis, its purpose is to discover the links that exist between the different types of data (we often talk about patterns). As part of the use of data analytics methods, it is assumed that there are correlations within the data and algorithms help us to find them and derive insights.

But in your choice of the algorithm, you must know that the algorithms are not all intended for the same uses. They are usually classified into two components: (i) the learning mode and (ii) the type of problem to be treated.

Table 1. Algorithm and their cases of uses

	Algorithm	Learning Mode	Problem to Be Treated
Simple	Simple regression	Supervised	Regression
	Multiple regression	Supervised	Regression
	Naïve Bayes	Supervised	Classification
	Logistic regression	Supervised	Classification
	Hierarchical Classification	Unsupervised	Clustering
	K-means	Unsupervised	Clustering
Complex	Decision tree	Supervised	Classification – Regression
	Random forest	Supervised	Classification – Regression
	Bootstrapping	Supervised	Classification - Regression
	SVM	Supervised	Classification - Regression
	Neural Networks	Supervised	Classification - Regression
	KNN	Supervised	Classification - Regression

In the following table you will find a classification of the most used algorithms to launch a big data project:

To launch a big data project, you have to master the way it works, in this context we propose two approaches:

Bottom-Up Approach

This approach goes from the bottom (the technique) to the top (the organization). With this approach, you will first validate the technical choices through a PoC and a case of use that you consider relevant. Do not immediately launch a major project that will transform your activities and put the data at the heart of your business issues. Once the project has been validated, you can continue with other experiments on ancillary domains (data analysis, visualization ...) or quickly realize a use case and bring value immediately. This organization is highly iterative both technically and functionally. This is obviously the method that brings the fastest results and can support your business strategies; in contrast, its visibility is limited.

Top-Down Approach

Completely the opposite of the previous method the top-down approach will first impact the organization of your company, transform it, to enable it to

launch big data projects. You will define a big data strategy for your entire business, a schedule of implementation of the concrete objectives that often result in new offers for the company or the improvement of existing offers. With this approach, the concrete results are longer to obtain. In contrast, objectives, responsibilities, and sponsors are clearly identified.

CONCLUSION

I remember my mathematics courses. We had to do "demonstrations". For example, demonstrate that point A is well located on the perimeter of circle C; it was much more complicated in general, it is just a simplified example. To begin this demonstration, it was necessary to start from an "idea". Thinking about that, I can see my teacher, who also said: "So who has ideas?" And fortunately I was among the geniuses of my class and I always arrived with a starting idea that, even though it seemed straight out of nowhere, leads me to the good conclusion. My classmate was always asking me, "How did you think about it?" You will understand that she was not very good at mathematics, she told me: for me, "the idea" was more in the realm of philosophy and art than mathematics.

Yes! Even in the world of geometry or algebra, the idea, this thought arising in one-knows-what-corner-of-the-brain, is what can save a situation. This can lead to an extraordinary communication campaign. That is what can hold a customer ... And that is true in any sector.

So, if you want to launch your big data project, you have to find the idea that will change the situation and add value to the business playground. You have to define your objectives behind this idea for what you have to work in order to success your project. But, do not forget that you need to collect, clean and prepare your data, because this data will be analyzed and converted into patters or model, or the result of your idea.

So, to success your big data project and carry out a process of data analytics using the diversity of algorithms, it is best to consider an algorithm as a recipe, data as ingredients, while the machine is like a mixer that supports a lot of the difficult tasks of an algorithm, in order to convert your idea into value.

REFERENCES

Davenport, T. H., & Kim, J. (2013). *Keeping Up with the Quants: Your Guide to Understanding and Using Analytics*. Harvard Business Review Press.

De Mauro, A., Greco, M., Grimaldi, M., & Nobili, G. (2016). Beyond Data Scientists: a Review of Big Data Skills and Job Families. *11th International Forum on Knowledge Assets Dynamics – IFKAD 2016, Towards a New Architecture of Knowledge: Big Data Culture and Creativity*.

Harrington, K. (2017). *Entrepreneurial ecosystem Momentum and maturity: The Important Role of Entrepreneur Development Organizations and Their Activities*. Ewing Marion Kauffman Foundation.

Sedkaoui, S. (2018a). *Data analytics and big data*. London: ISTE. doi:10.1002/9781119528043

Sedkaoui, S. (2018c). How data analytics is changing entrepreneurial opportunities? *International Journal of Innovation Science*, *10*(2), 274–294. doi:10.1108/IJIS-09-2017-0092

Shroff, G. (2013). *The Intelligent Web, Search, Smart Algorithms and Big Data*. Oxford, UK: Oxford Univ. Press.

Siegel, E. (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die*. John Wiley & Sons.

Siegel, E. (2016). *Predictive Analytics*. New York: Wiley.

KEY TERMS AND DEFINITIONS

Business Model: A business model is a company's plan for how it will generate revenues and make a profit. It explains what products or services the business plans to manufacture and market, and how it plans to do so, including what expenses it will incur.

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 Lake: Is a collection of storage instances of various data assets added to the originating data sources. These assets are stored in a near-exact, or even exact, a copy of the source format. The purpose of a data lake is to present an unrefined view of data to only the most highly skilled analysts, to help them explore their data refinement and analysis techniques independent of any of the system-of-record compromises that may exist in a traditional analytic data store (such as a data mart or data warehouse).

Data Mining: This practice consists of extracting information from data as the objective of drawing knowledge from large quantities of data through automatic or semi-automatic 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.

Entrepreneur: Entrepreneurship is not only an outcome of the ecosystem but also an important input factor since entrepreneurs drive the ecosystem by creating it and keeping it healthy. Drucker believes that what entrepreneurs have in common is not personality traits but a commitment to innovation. For innovation, the entrepreneur must have not only talent, ingenuity, and knowledge but he must also be hardworking, focused and purposeful.

Missing Values: Occur when no data value is stored for the variable in an observation.

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

Outliers: An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

Small Business: Are companies that fall under specific legal limitations regarding the number of employees and the annual turnover. However, this differs from one country to another.

Chapter 9
Big Data Analytics in Action: Examples

ABSTRACT

Have you ever wondered how companies that adopt big data and analytics have generated value? Which algorithm are they using for which situation? And what was the result? These points will be discussed in this chapter in order to highlight the importance of big data analytics. To this end, and in order to give a quick introduction to what is being done in data analytics applications and to trigger the reader's interest, the author introduces some applications examples. This will allow you, in more detail, to gain more insight into the types and uses of algorithms for data analysis. So, enjoy the examples.

INTRODUCTION

The secret of getting ahead is getting started. The secret of getting started is breaking your complex overwhelming tasks into small manageable tasks, and starting on the first one. (Mark Twain)

If you ask a data scientist about which algorithm is best for analyzing such a problem, he will ask you to try several and see which one works best depending on your case (Sedkaoui, 2018a).

DOI: 10.4018/978-1-5225-7609-9.ch009

So, it depends on:

- The data quality
- Parameters that will be used
- Data source
- Execution time required
- Available parameters to influence the performance of the algorithm
- Etc.

So, for each type of analytics question, a specific group of methods called algorithms. The algorithms present a sort of real evolution. They make our programs smarter by allowing them to learn automatically from the data we provide. We noticed this importance again during a participation in several events, meets and conferences around the world during these last few years. Personally, I was able to perform even better, thanks to several applications, how do the algorithms work? And how value id generated?

With the several examples illustrated in this last chapter, you can see and experience how data analytics will generate added knowledge and how it turns ideas into business opportunities. Big data analytics revolutionizes businesses by mixing 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.

The aim of these examples is to show you how you can read your data, apply tools and methods and visualize your results. Going from the simple example, the aim is then to help you learn to practice in this universe.

Far from the big abstract speeches, the author will make you, through this chapter; discover the practices of data scientist. And it will be the opportunity for you also to put your hand in this field, with just enough theory to understand what involves the methods of data analysis used, but especially with your computer, some free and powerful big data software and technologies, as well as a little thought, you will participate actively in this passionate exploration

HOW CAN SMALL BUSINESS USE BIG DATA? PRACTICAL EXAMPLES AND APPLICATION CASES

Through the collection, analysis and the data exploration, big data is synonymous with innovation in term of use and contribution in the enhancement of competitive advantage of your business activity. Applications taking advantage of big data are announced as numerous, diverse and very promising. Reading the press, we hear often about the recommendation systems of US giants like Google and Amazon, decryption of the human genome, monitoring logs or behaviors, retargeting, etc.

Davenport and Kim (2013) and Brynjolfson and McAfee (2012) and others suggest that large companies have used big data analytics to increase their performance, and some uses can be extracted:

- Processing a large amount of data generated through internal operational systems and processes.
- Capturing external sources of data and using them to enrich the data generated internally.
- Scanning the external environment and industry for information on what is being said by competitors and customers, using technologies such as sentiment analysis

Historically, analytics were used to perform simple counts and analyze frequencies. Then data mining began to relate different phenomena while taking into account larger volumes of data, but mainly accessible and structured. Ultimately, 'data science' is inevitable as it can help extract various kinds of knowledge from data. The broader term data science, which can be applied to many types and kinds of data big and small, captures a theoretical and methodological sea change occurring in educational and social science research methods that is situated apart from or perhaps between traditional qualitative and quantitative methods (Gibson and Ifenthaler 2017; Gibson and Webb 2015).

Today, using machine learning algorithms, data analytics makes it possible to exploit all types of data to implement new models of interpretation and prediction, to derive forecasts on future developments and make recommendations on the actions to be taken. The most common examples:

Predictive Maintenance

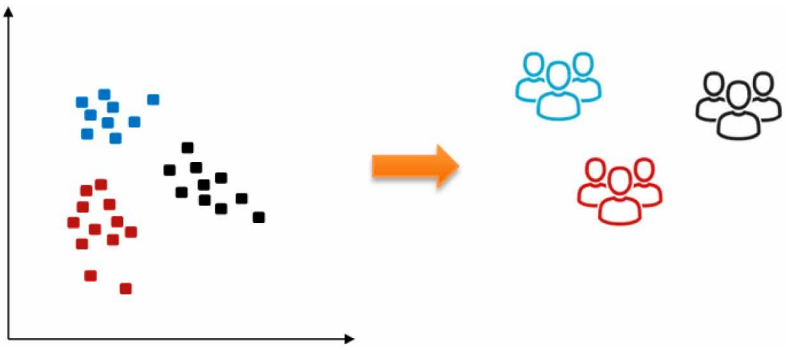
Analyze the signals in the production line to anticipate failures and change parts before the failure occurs. But, the integration of big data for the purpose of predictive maintenance is the result of careful thought and a complex process. It is a question of seizing the mechanisms occurring at the time of the breakdowns, and thus to establish an "alert" about the problem. To this end, dozens of sensors are connected to the production line measuring the various relevant signals relating to the proper functioning of the chain. And for that we must distinguish three stages:

- **Understand:** This is where the company's strategy comes in. Monitoring all equipment can be too costly and irrelevant. The company will, therefore, have to prioritize its actions in order to identify the strategic processes on which the optimization must be carried out, for example, rationalization of a production and distribution chain, and to understand the key parameters involved in the phenomenon. Once the intervention points have been decided, the data collection is launched, supported by the support of a Master Data Management
- **Alert:** The analysis of the received information will then make it possible to predict more or less complex events. Starting from the tide of data provided, several alerts will emerge. Alerts will then be brought to the attention of decision-makers.
- **Integrate:** Once informed, the decision-maker will be able to integrate a new understanding of its tools to adapt both its strategy and cost control. By following this logic, the company will be able to use the data provided to move from a reactive approach to a proactive approach through predictive maintenance.

For example, in 2015, Air France KLM used MongoDB to analyze and anticipate the breakdowns of its Airbus A380s. With the help of 300000 sensors placed at RCG airport, the flight data of the aircraft are transmitted to the Air France KLM engineering and maintenance center for analyzing and detect any weak signal announcing a future failure. In less than one hour, the diagnosis is delivered to the maintenance teams present at the airport and available in case of intervention.

The solution set up by MongoDB allows the company to detect upcoming faults 10 to 20 days before they occur, the time to identify and locate the

Figure 1. Customer Segmentation example



source of the fault has passed on average from 6 hours to 5 minutes and 75% of breakdowns were avoided.

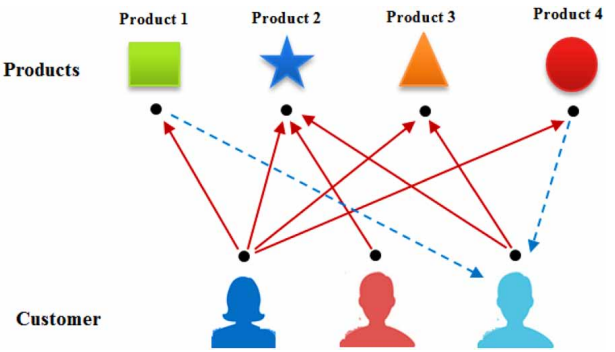
Customer Segmentation

Customer segmentation is a key element of marketing, particularly in the e-commerce sector. It allows groups of consumers with similar behaviors to be grouped together in various categories and to analyze these behaviors, then optimize product recommendations, marketing campaigns, and thus the sales of the company.

This customer segmentation can be done within a data management platform (DMP) or with specific algorithms adapted to the customer. The general data analytics model used is called Clustering, which is an example of an unsupervised algorithm, unlike the models: Classification and Regression that we have seen previously.

In the multi-dimensional space of all consumer characteristics; for example, the frequency of purchases, the average price of purchases, responsiveness to recommended products, demographic information such as location, age, gender ... potentially hundreds of characteristics; each consumer is represented by a vector. A distance adapted to the desired type of segmentation then makes it possible to group the 'near' vectors, to obtain categories of consumers and to analyze all their common characteristics.

Figure 2. How it works



Customer Behavior Analysis

Understand the customer motivations, identify the best ways to prospect and retain them to offer each of the personalized services at the right time by the most relevant channel, is the main objectives of many companies. Other data analytics techniques are used to analyze customer behavior in the e-commerce sector, luxury goods, supermarkets ... as well as social networks like LinkedIn, Facebook Twitter, etc. These advanced techniques are called collaborative filtering or matrix factorization. As an illustration, Figure 2 describes how it works in the case of a product recommendation.

In a so-called co-occurrence matrix, we identify all products pairs that are frequently purchased together. Then, by a statistical algorithm, we deduce a distance between products encoding this craving to be purchased together. The algorithm then determines, for each consumer who bought a particular product, what would be the most appropriate product to recommend him, that is to say, that minimizing the distance with the product already purchased.

Fraud Detection

Fraud, as we know, is a very promising field of application for big data. For a client, the company can realize the design of an automatic fraud detection system based on accident data for example, and thus launched alerts on claims files, policyholders, sales networks, benefits of statistically abnormal behavior…

Detect fraud or failure to take action before it happens. In this case, algorithms can detect fraudulent and abnormal behavior. These algorithms

are widely used to detect financial frauds. Here is a simplified example: if we consider that the average price of a medicament is 25 dollars, and if we cross this information to the profile of a consumer who has used his credit card with his socio-demographic information, it is easy to identify for example, that the purchase of a 400 dollars is questionable. This is especially true if it was made 60 minutes ago and 200 km from the previous transaction.

Also, it is unlikely that an individual from Spain for example, who spends 800 € monthly, spends a sudden 10000 € in Ukraine. The algorithm will signal this action as potentially a fraud.

These opportunities are being sold as part of a data revolution being labeled big data, and companies are taking advantage of this revolution to market new products and services to institutional leaders who know they should be harnessing this big data but are not entirely sure how to go about doing so (Lane, 2014; Sedkaoui, 2018c).

Monetize data volumes, stored in large databases, and derive value from them, is one of the best-known goals of big data. But the potential of the techniques, methodologies, and examples that go into the definition of data analysis goes far beyond simple data value. In what follows you will discover what you can do to promote your entrepreneurial activities or your business, even online, through analytics.

Several techniques and practical examples selected so that everyone can better know this vast field, in continuous evolution. By using the techniques and taking inspiration from the following examples, you will be able to understand, develop and refine your business strategy and distinguish yourself from your competitors.

So let's do something smart with the data! Here is a list of examples to start today in the field of marketing and more. Do not expect a revolution, the application cases revolve around the daily topics of the business activities: acquisition, retention, conversion and customer knowledge.

Cluster Analysis to Identify Target Groups

Cluster analysis makes it possible to identify a group of users (in databases) according to common characteristics. These characteristics can be age, geographical location, occupation, and so on. This is a data analytics technique that is used in marketing to segment the database and to send, for example, some promotion to the right target for a particular product or service (young,

retirees, etc.). The combinations of variables are infinite and make the cluster analysis more or less selective depending on the search requirements.

Regression Analysis for Prediction

Predicting the future is the dream of any business professional. Without the need for a crystal ball, we benefit from regression analysis, a data analysis technique that allows us to study changes, habits, customer satisfaction levels and other factors related to parameters such as budget of an advertising campaign ... Once you change any of these settings, you will have a pretty close idea of what will happen to your user audience.

Classification Analysis to Identify Spam

How to classify a customer's reply email? And how identifying potential correlations between potential buyers of your products before and after the implementation of an advertising campaign? The answer is unique: the classification analysis, the analytics technique that allows recognizing patterns (recurring patterns) within a database. It is an effective way to make your business strategy more efficient, eliminate redundancy and create optimized sub-archives.

Anomaly Detection to Recognize Inconsistencies

Every business, whether big or small, has to deal every day with the consequences of possible mistakes made by top management, employees, suppliers or customers. A common mistake in the data entry phase or when buying a product has the same effect as a pebble in a shoe. Nothing upsetting, but it is still embarrassing. In order to eliminate database inconsistencies and anomalies at the source, a special data analytics algorithm is used that is called anomaly detection. In this case, again, it will be our software to manage the search, which is programmed to perform complex operations on databases that may contain hundreds of thousands of records (addresses, names, etc.).

Intrusion Detection for Better Security of the System

Marketing and security are two aspects that seem to have no connection and should be linked. Think about the harmful effects that a Direct Email

Marketing (DEM) campaign could have on a contaminated database. To avoid using intruder-infected archives; values added by crackers or real viruses that duplicate data; it is enough to search for intruders, a technique that cleans up the database and guarantees greater security to the entire system.

Decision Trees to Optimize Risk Management

Whenever you make a decision, you are at a crossroads. When there are many options, you find a decision tree instead of crossing. At first, glance, dealing with a tree like this confuses ideas, but if we have sophisticated IT tools that organize the tree and submits definitive choices, including costs/benefits, everything changes, and the tree becomes a valuable tool for Project Risk Management. Here again, the depth of analysis depends largely on the technology available: the more advanced the software, the more the tree will show you the best way to follow.

Neural Networks to Automate Learning

The artificial neural network concept is complementary to clustering and decision trees. This is one of the newest data analytics applications, based on which the machine you use for your business actions, and thus the computer that manages your database learns to identify a certain pattern within which are present elements having precise relations between them. The result of this learning is the recognition and memorization of patterns that may be useful, not necessarily immediately but in the future, to decide if the goal is achieved and how. This algorithm can help you know more precisely the composition of the target of a product or service.

Association Rules to Analyze the Relationships Between Data

The typical use of the association rules concerns sales of products, especially for large volumes. Whether online through an e-commerce website or in person at a market, there may be interesting relationships between the data you have. Relationships you did not suspect or even imagined. For example, 90% of customers who buy an online product also buy another, always the same (the case of Walmart). These details allow us to create targeted marketing offers, special promotions and winning formulas.

Data Storage for Big Data Processing

The last essential data analysis technique, which may be more correct to call an application, is called data warehousing. Here we enter the field of customer profiling, especially with regards to big data processing. Choose software, like Egon for example, for data warehouse means simplifying databases, extrapolating the most interesting information about your customers, facilitating the creation of detailed reports and more.

A VERY SIMPLE EXAMPLE OF DATA ANALYSIS ALGORITHMS: TITANIC SURVIVORS

Even if you need to acquire a good knowledge of statistics and mathematics in order to carry out your big data project and better guide its success, as well as certain other skills in various programming languages, your approach is within everyone's reach. When dealing with data analytics problem, it is important to establish a first benchmark to set a "first level" of how algorithms work? From this level, you can then enjoy the use of other algorithms and you will appreciate making big data into work.

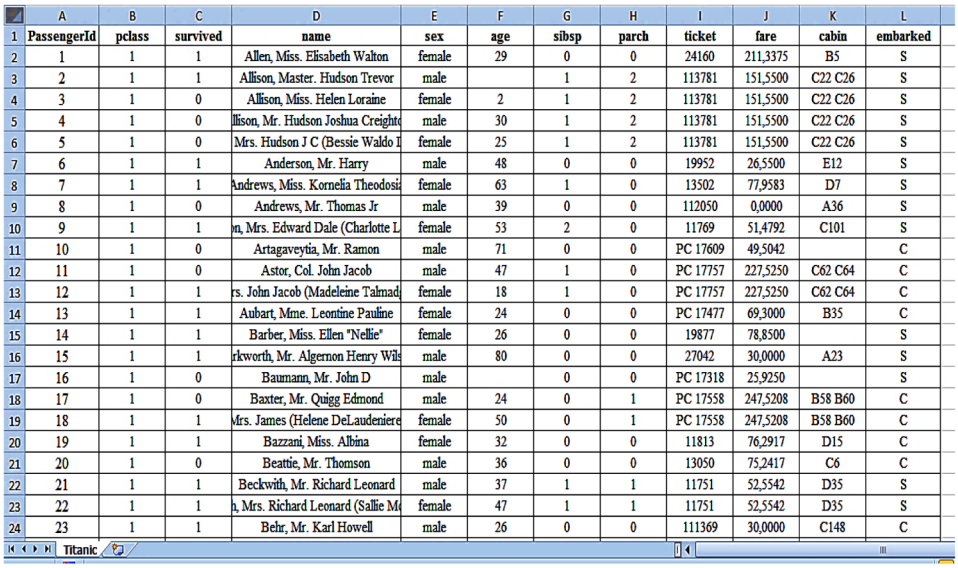
To clarify the process, let's look together how we can exploit data and derive insights by a very simple example of an algorithm that finds the relationships and correlations between data to predict events, a classic example that we often encounter: How to predict the survivors of Titanic?

The dataset is taken from the machine learning competition Titanic: Machine learning from Disaster on "*Kaggle*", the famous data science platform.

We will use a simple dataset about the historical data of Titanic passengers. We create an Excel table (in csv format) with online per passenger and its characteristics in the column. If the passenger survived we put 1 if it is not the case we put 0. We adopt the same coding procedure for other variables such as sex (1 for men and 0 for women); the class (ticket) respectively: 1, 2 or 3; in addition to other information about: the passenger age; the number of his family members on board and the price of his ticket in dollars, as illustrated in the figure below.

The task is to create an algorithm that allows us to predict the first column from the other columns to this is to say we want to classify Titanic passengers into two groups: survivors and no-survivors, and categorize them according to the other characteristics mentioned in Figure 3.

Figure 3. Titanic survivors' databases



	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked
2	1	1	1	Allen, Miss. Elisabeth Walton	female	29	0	0	24160	211,3375	B5	S
3	2	1	1	Allison, Master. Hudson Trevor	male		1	2	113781	151,5500	C22 C26	S
4	3	1	0	Allison, Miss. Helen Loraine	female	2	1	2	113781	151,5500	C22 C26	S
5	4	1	0	llison, Mr. Hudson Joshua Creight	male	30	1	2	113781	151,5500	C22 C26	S
6	5	1	0	Mrs. Hudson J C (Bessie Waldo I	female	25	1	2	113781	151,5500	C22 C26	S
7	6	1	1	Anderson, Mr. Harry	male	48	0	0	19952	26,5500	E12	S
8	7	1	1	Andrews, Miss. Kornelia Theodosi	female	63	1	0	13502	77,9583	D7	S
9	8	1	0	Andrews, Mr. Thomas Jr	male	39	0	0	112050	0,0000	A36	S
10	9	1	1	n, Mrs. Edward Dale (Charlotte L	female	53	2	0	11769	51,4792	C101	S
11	10	1	0	Artagaveytia, Mr. Ramon	male	71	0	0	PC 17609	49,5042		C
12	11	1	0	Astor, Col. John Jacob	male	47	1	0	PC 17757	227,5250	C62 C64	C
13	12	1	1	rs. John Jacob (Madeleine Talmad	female	18	1	0	PC 17757	227,5250	C62 C64	C
14	13	1	1	Aubart, Mme. Leontine Pauline	female	24	0	0	PC 17477	69,3000	B35	C
15	14	1	1	Barber, Miss. Ellen "Nellie"	female	26	0	0	19877	78,8500		S
16	15	1	1	rkworth, Mr. Algernon Henry Wils	male	80	0	0	27042	30,0000	A23	S
17	16	1	0	Baumann, Mr. John D	male		0	0	PC 17318	25,9250		S
18	17	1	0	Baxter, Mr. Quigg Edmond	male	24	0	1	PC 17558	247,5208	B58 B60	C
19	18	1	1	Mrs. James (Helene DeLaudeniere	female	50	0	1	PC 17558	247,5208	B58 B60	C
20	19	1	1	Bazzani, Miss. Albina	female	32	0	0	11813	76,2917	D15	C
21	20	1	0	Beattie, Mr. Thomson	male	36	0	0	13050	75,2417	C6	C
22	21	1	1	Beckwith, Mr. Richard Leonard	male	37	1	1	11751	52,5542	D35	S
23	22	1	1	h, Mrs. Richard Leonard (Sallie M	female	47	1	1	11751	52,5542	D35	S
24	23	1	1	Behr, Mr. Karl Howell	male	26	0	0	111369	30,0000	C148	C

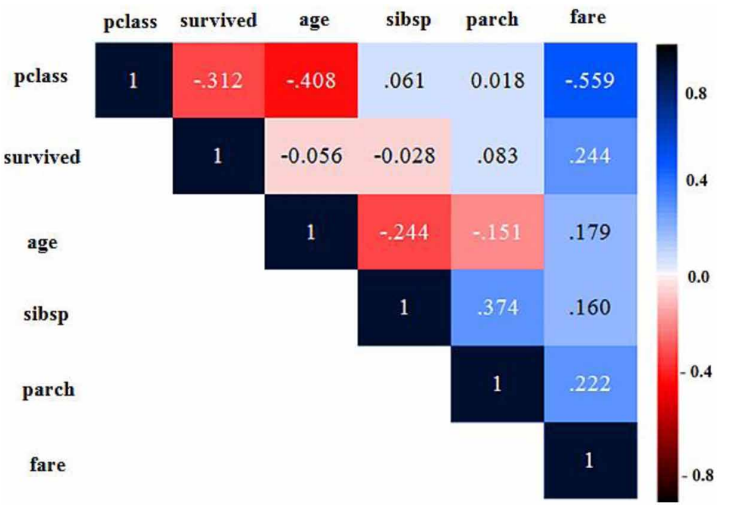
Titanic

The algorithm will learn, for each passenger in the database, the relationships between the different characteristics on the passenger and the fact that he survived or not. Then, given the information of a new passenger, the algorithm can predict whether it will survive or not.

Look first at the data structure; the columns contain the following information:

- **PassengerId:** Number of the passenger
- **Pclass:** The class in which the passenger traveled (1 for the first class, 2 for the second class, or 3: for the third class)
- **Survived:** Indicates the death or survival of the passage. This is the vector to predict for the test. The result is Boolean: 1 if the passenger survived, otherwise 0
- **Name:** The name of each passenger, therefore categorical variables
- **Sex:** Male or female
- **Age:** By years
- **SibSp:** Number of spouses and siblings on board
- **Parch:** Number of parents/children on board
- **Ticket:** Ticket reference
- **Fare:** Price paid
- **Cabin:** Cabin number (for those who have one)

Figure 4. Correlation matrix



- **Embarked:** Port of Departure (S = Southampton, C = Cherbourg, Q = Queenstown)

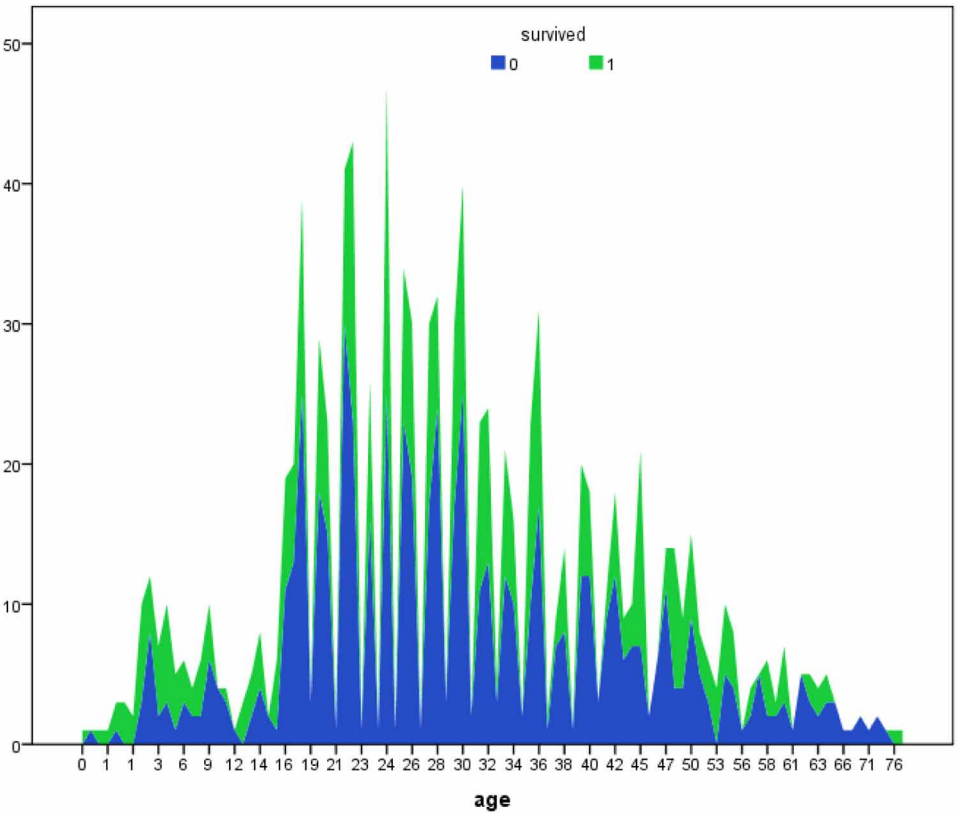
Before starting the manipulation of the algorithm, we must go through a very important phase which consists of analyzing and processing the data to better prepare the learning dataset. A first dataset analysis allows us to distinguish three types of variables:

- **Continuous Variable:** Age, Sibsp, Parch, Fare.
- **Categorical Variable:** The values of these variables are part of a well-defined subset: Pclass, Sex, embarked.
- **Text Variable:** Name, ticket. This type of variables are little or hardly exploitable, but it is possible to extract useful information.

In order to determine which variable are relevant for the analysis, we will look at the variables for which we can plot the correlation matrix (Figure 4).

I want by this image show you how we can create a visual graph which helps us to better read our data. The analysis of this image allows us to determine the most relevant variables to integrate into the model. Survival is negatively correlated with class (pclass); this is to say that passengers of the third class have survived less, and positively with the price ticket (fare).

Figure 5. Chance of survival by age category



The age and size of the families also, but to a lesser extent, influence the chances of survival.

Figure 6. Survivors by sex

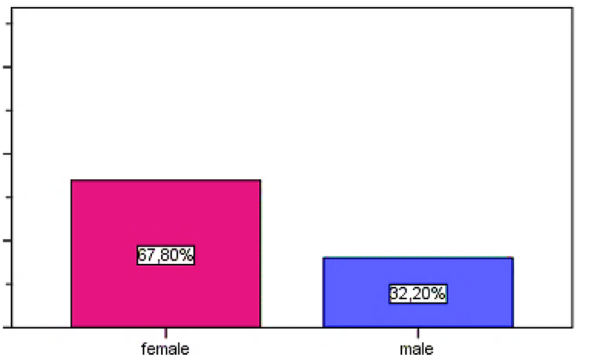
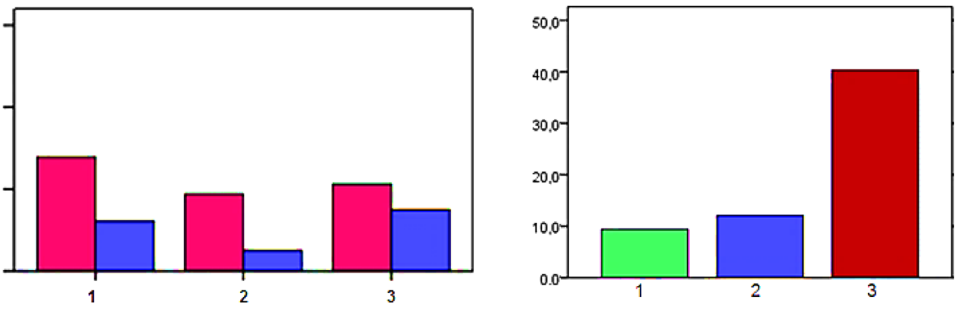


Figure 7. Survivors by class passenger



Regardless of the sex of the passenger, Figure 5 shows that the chances of survival are higher for young people (less than 20 years old) and for the elderly.

Now let's analyze categorical variables. When we watched the Titanic movie, for the first time we remember this sentence: "Women and children first!". We can assume that the chances of survival are higher for women, so we will analyze this supposition and see then what is the result?

Figure 6 confirms this intuition: only 32% of men have survived against more than 67% of women.

Figure 7 shows the survival rate is much higher for first-class passengers (more than 40%) than for third-class passengers. We can say also that there is a significant disparity between men and women regardless of class.

Also, in order to be able to establish a more precise model, it is necessary to treat the missing value or outliers.

As we check our database we notice that there are missing values only for two variables: the age (263 missing value) variables and the cabin number that corresponds to passengers who have traveled in a Spartan way, so it is not really a missing value. To treat these data I will proceed as follows:

- For the age, I assume that first class passengers are mostly older and richer than those in the third class. So I will sort by class and assign to the missing values the median age of the passengers of the corresponding class.
- For cabins, I will establish a simple model and assign 0 if the passenger has no cabin or 1 if he has one (regardless of the class).

Once our database is ready, we can start and move on to the modeling phase, the goal of this phase is to find the model that can predict the output

Figure 8. Decision tree of Titanic survivors

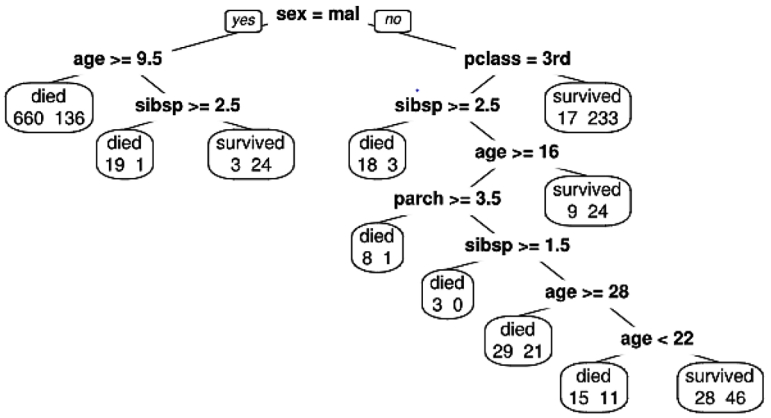


Figure 9. Decision tree results application

	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked
821	820	3	1	Glynn, Miss. Mary Agatha	female		0	0	335677	7,7500		Q
822	821	3	1	ith, Master. Frank John William "F	male	9	0	2	363291	20,5250		S
823	822	3	0	Goldsmith, Mr. Frank John	male	33	1	1	363291	20,5250		S
824	823	3	0	Goldsmith, Mr. Nathan	male	41	0	0	ON/O.Q. 310	7,8500		S
825	824	3	1	ith, Mrs. Frank John (Emily Alice	female	31	1	1	363291	20,5250		S

value in the case of our example. There are about ten algorithms applicable to this problem, but we chose two types of algorithms which will help you put your hands in the dough: the decision trees and the random forest.

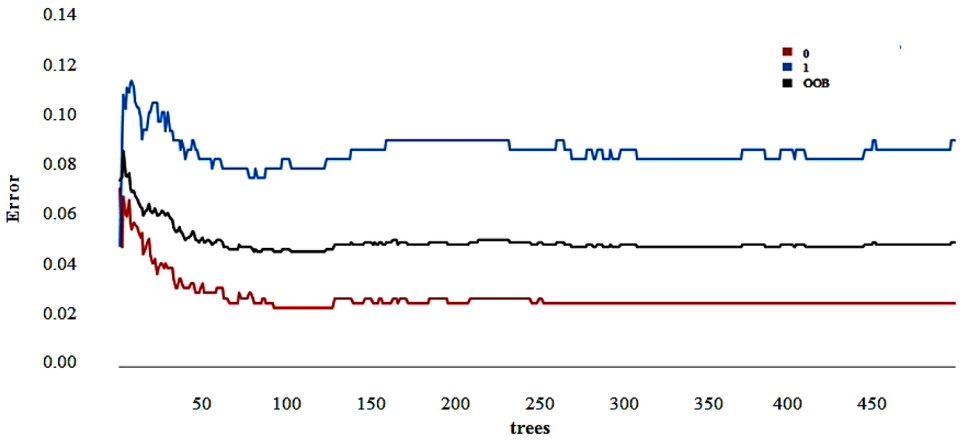
Decision Tree

Figure 8 shows the result of the execution of the decision tree algorithm on our Titanic passenger database, using R (the function used is "Rpart").

Now, taking the example of the 822 passengers and we will try to predict whether this passenger survived or not. His name is Mr. Frank John Goldsmith, he is a man, we are on the left side of the first knot, he is 33 years old, then over than 9.5, we are sailing on the right side, he was in 3rd class, and he had only one family member on board, so less than 2.5 we are now on the left side, at the end we deduce that Mr. Frank died at the sinking of the Titanic, and that is how a prediction is made, using the decision trees (Figure 9).

The question that maybe you are asking now is: how do you choose the criteria and the division values? Why do we start with the passenger's sex and not with another variable? It is the intelligence of the algorithm that does it for

Figure 10. Algorithm performance



you, the principle is simple: we must start with the most discriminating variables to the least discriminating one, that is to say, we must create the most disorder possible, and this principle is modeled mathematically by the entropy function.

Random Forest

The advantage of the Random Forest algorithm is the independence between the trees to be trained, which makes it possible to parallelize the treatments and improve the performance of the algorithm. Each tree will be different from the others and their grouping should allow presenting less variance than a single tree. To build my model, I selected the parameter. To evaluate the performance of the model, we precede a cross-validate.

This corresponds to the error OOB made by taking into account all the trees. In addition to the OOB error, we also have a column by category representing the evolution of the error made on each of the modalities of the response variable. The graph (in Figure 10) is the result of running the Random Forest algorithm.

It is found that the error stabilizes at 5% after 300 pieces of training (training of 300 under trees randomly drawn). It is important to specify a good number of trees (also call number of iterations) for the algorithm, an underestimation of the number of iterations increases the risk of errors, an overestimation of the number of iterations impacts the performances of the algorithms (time required to execute the algorithm).

This is a crucial passage for anyone who wants to work with big data. It is a very easy dataset, even if you think that predicting the survivors of Titanic is a problem without any interest for you, but this example allows you to apply the different methods and algorithms discovered in previous chapters. Indeed, it is important to see how you can understand your data, to know how to manage the missing variables, to establish an explanatory model, to realize a tree of decision ... that will allow you to understand and to constitute your own manual of instructions when working with data.

USE HADOOP MAPREDUCE FOR YOUR ANALYTICAL PROJECTS

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 (Walwei, 2016). Today, we hear a lot about big data technologies: the project managers talk about it and want to experiment with the contribution of these technologies in terms of scalability, businesses talk about big data and DataLab missions, HR are looking for big data experts ... A large number of big data frameworks has therefore emerged in recent years because the big data ecosystem is booming.

In this part of the book, we will return again to the big data context in which it is of interest because it allows the scaling of processes on large volumes of data. However, in order to better guide your big data project, it is required to associate it with a dedicated software infrastructure that allows the analysis to be executed in a massively distributed way on a cluster of machines while taking charge of distributed computing issues.

Many companies have been using Hadoop for a long time to analyze big data. This framework is based on simple programming models to ensure data processing and make them available on local machines.

Hadoop is an open source framework for the development of distributed scalable applications. It allows the distributed processing of large volumes of data on a cluster of several hundred (or thousands) of standard machines called commodity hardware. It is based on (i) HDFS (Hadoop Distributed File System) for the data storage part, and (ii) MapReduce, which is a programming model (with an associated implementation) for parallel processing and distributed data on a cluster.

HDFS is the foundation of Hadoop since it is responsible for storage. It is a distributed file system whose management is facilitated by the framework that it integrates. The latter, for example, includes commands familiar to Linux users, such as *ls*, *mkdir*, *rm* or *cp*. The only difference is that these commands must be preceded by '*hdfs dfs'*. Example of an order: "*hdfs dfs-ls /data/financial/*".

These features make it possible to parallelize operations on the data, to be fault resistant (thanks to replication) and to be easily scalable. This is the advantage of HDFS over a classic RDBMS. HDFS consists of two types of components:

- **The *NameNode*:** This is the "master" component of HDFS, which controls the distribution of data and their location;
- **The *DataNode*:** Responsible for storing data on its own machine and replication on other DataNode.

MapReduce is a programming model and a software framework that you can use to create data processing applications. Originally developed by Google, MapReduce allows fast and parallel processing of large datasets on node clusters. This framework has two main phases: the *map* phase to separate the data to be processed, and the *reduce* phase to analyze the data.

- **Map:** The data is extracted by the mappers, transformed and then prepared for the reducers.
- **Reduce:** Data is collected from mappers and processed and analyzed.

It may be more technical for you! Okay let's get away from IT context and a big data universe for a few moments; this gives us a little time to understand this framework. Let's talking about orange juice, I think that you know how to prepare orange juice? If so, you will understand the two phases of the MapReduce.

- **Peeling Orange:** Orange must be peeled one by one;
- **Pressing Orange:** We put all the oranges in a press, and we have our juice.

It summarizes the process. You have a pile of oranges; you peel them one by one. When all the oranges are peeled (and not before!), you can start the second phase. As our goal is to translate a problem into a model that we will

call MapReduce, we distinguish precisely the boundary between the two phases, and the tasks performed on each side.

Firstly, we have the transformation workshop: it is the peeler, who takes an orange peel it or prepares it for the second phase. Secondly, we have the assembly workshop: where we put a pile of peeled oranges in order to produce orange juice.

We can already draw two lessons on the essential characteristics of our basic process. The first is the transformation workshop that applies an individual operation to each product. The second is about the assembly workshop which, on the contrary, applies a transformation to the grouped products.

I guess that you have probably a simple idea concerning this framework. Now, it is time to take a little height and characterize the MapReduce model (see Appendix) in computer terms.

The principle of MapReduce is old and it is summarized as follows: given a collection of items, we apply to each item an individual transformation process: or the map phase, that produces intermediate values. These values are grouped by the label and subjected to an assembly function applied to each group: reduce phase. The Map phase corresponds to our transformation workshop, the Reduce phase represents our assembly workshop.

Let's take the model in detail.

Input Item Concept

An input item is any value that can be subjected to the transformation function. In our culinary example, the entry items are the raw fruits: apples, oranges, pineapples, etc. The transformation applied to the items is represented by a Map function.

Map Function Concept

The Map function is applied to each item of the collection and placed in an accumulator. In our example, this function is peeling. For the same fruit, several values are produced, or none if the fruit is rotten. It is often necessary to partition the values produced by the map into several groups. Just modify the map function to not emit a value either, but associate each value with the group to which it belongs.

The map function produces, for each item, a pair $\left(k, v\right)$, where k is the identifier of the group and v refers to the value to be placed in the group.

The identifier of the group is determined from the item processed (this is what we informally called label). In the MapReduce model, the intermediate data is the data produced by the Map phase.

Intermediate Pair Concept

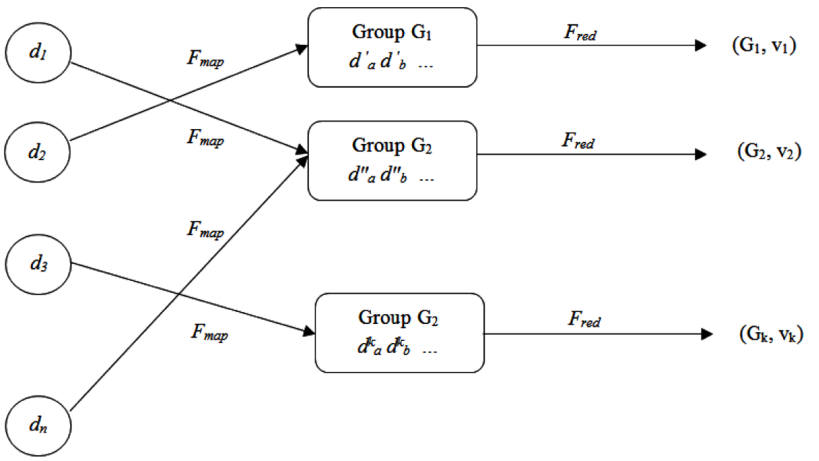
An intermediate pair is produced by the Map function; it takes the form (k, v) where k is the identifier (or key) of a group and v the value extracted from the input item by map function. For our example, there are many groups, and therefore many possible identifiers: apple, orange, pineapple ….

At the end of the Map phase, we have a set of intermediate pairs. Each pair is characterized by the identifier of a group, one can constitute the groups by grouping on the value of the identifier. Intermediate groups are obtained.

Intermediate Group Concept

An intermediate group is the set of intermediate values associated with the same key value. So we will have the apple group, the orange group, and the pineapple group and so on. We then enter the second phase, called Reduce. The transformation applied to each group is defined by the Reduce function.

Figure 11. Process summarized



Reduce Function Concept

The Reduce function is applied to each intermediate group and produces a final value. The set of final values (one for each group) is the result of MapReduce processing.

Figure 11 summarizes the MapReduce process. Let's now put it in the context of our documentary bases. We have a collection of data: $d1, d2, \cdots, dn$.

The map function produces intermediate pairs in the form of data dji, where $\left(j\right)$ designates the group of members. Note that: data input can generate several documents out of the map. Map function places each dji in a group Gj , $j \in \left[1, k\right]$.

When the map phase is over (and not before!), we can move to the reduce phase, which successively applies '*red'* function to the data in each group. For each group, we obtain a value vj .

Let's leave the world of fruit juice while and recall the good old paradigms of algorithm design, including the famous Divide and Conquer and its three stages, for a given initial problem, to:

- **Divide:** Split the initial problem into sub-problems;
- **Rule:** Solve sub-problems independently either recursively or directly if they are small;
- **Combine:** Construct the solution of the initial problem by combining the solutions of the different sub-problems.

MapReduce is divide to distribute to reign, in this sense the strategy is set up to perform a massive data calculation consists of splitting the data into smaller subsets, and assigning each set to a cluster machine thus allowing their parallel processing. It will then suffice to aggregate all the intermediate results obtained for each set to build the final result. In another way, the map phase independently associates the value 1 with each data item.

Figure 12. Application example

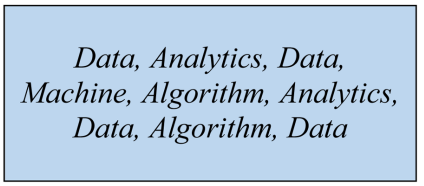


Figure 13. Split the input

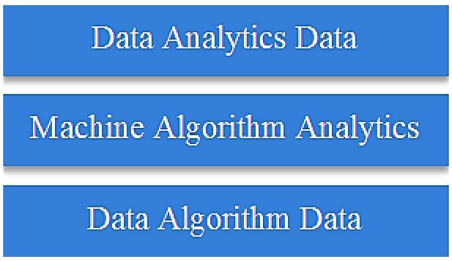
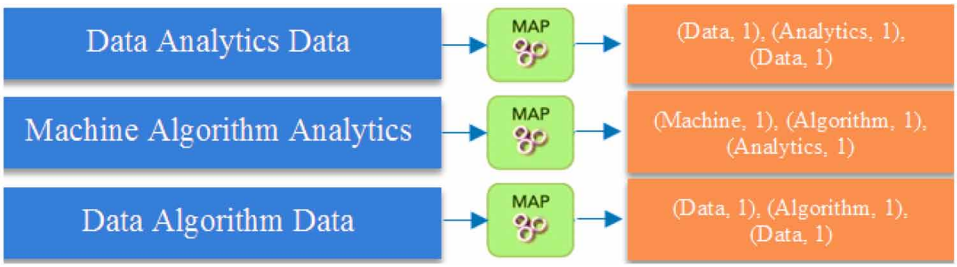


Figure 14. First phase: Map



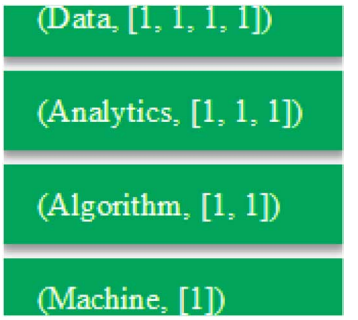
The reduce phase makes the reduction of the sum on each intermediate value. We count the number of data! For example, the key could be age, name, or first name, with corresponding values for each person in the database.

To make the MapReduce process more concrete, we will illustrate it with the "WordCount!" or the typical example of MapReduce, so typical that it has become the "Hello World!" of MapReduce and distributed computing. Nothing complicated. Let's take a collection of text documents as input: the goal of Wordcount is to calculate the number of occurrences of each word in the collection.

We will use here a collection of elements of type "(key, value)". The key corresponds to the word and the value corresponds to its number of occurrences. Suppose you can store your entire collection in one file; so it is really not a difficult problem. But, if your collection is very big it becomes more complicated and it is necessary to perform this count in a distributed way. This is where MapReduce comes in.

We will work on the following words, and calculate the number of appearance of each word (Figure 12).

Figure 15. List of pairs: Key, Value



We are far from billions of words available for example on Wikipedia but the goal is to illustrate the principle of MapReduce.

First Phase Map: A Processor per Word w That Reads the Text and Returns a Pair $(w,1)$

We will, therefore, assume that our input data has been split into different words.

We now need to determine the key to use for the map operation. The way in which we have responded to the problem sequentially directs us quite naturally to the choice of taking as keys the words of the text.

The next step is to determine the code of the map operation according to the schema imposed by MapReduce, that is to say, it must return a list of pairs (key, value). In the case of WordCount, the map operation will thus break down the text of the words provided as input and it will generate for each word a pair (word, 1).

So, now we have everything we need for the Map step, which is to apply the map operation to each word in parallel as shown in Figure 14.

At the end of the MAP step, we have several lists of pairs (key, value). We will consider that we are able to group and sort, by the common key, the intermediate results provided by the map phase. This corresponds to the Shuffle phase. This phase is fully managed by the MapReduce execution framework.

We now have at our disposal a set of pairs (key, list-of-values).

Figure 16. The overall MapReduce WordCount process

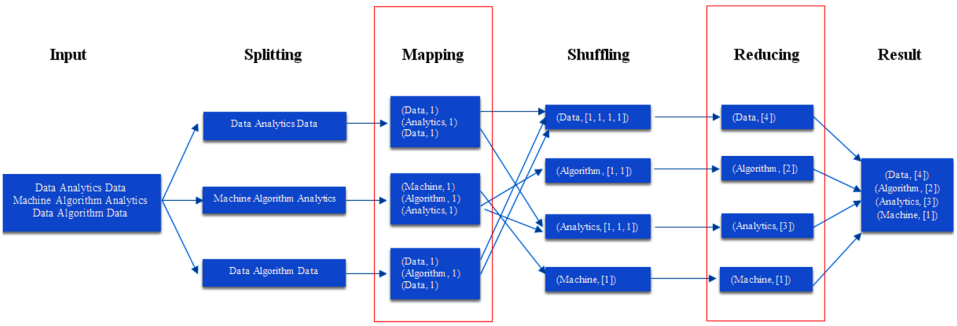


Table 1. Cars by model and price

ID	Model	Price
1	Car1	4200
2	Car2	8850
3	Car3	4300
4	Car4	3140

Second Phase Reduce: Reduction of the Addition for Each Word w, of the Set of Pairs $(w,1)$, Thus Calculate the Number of Occurrence of Each Word w in the Text

We now have to write the code for the reduce operation, according to the scheme imposed by MapReduce. For WordCount, the reduce operation will therefore just consist of summing all values in the list associated with a key.

The Reduce phase can, therefore, be applied. It consists of applying the reduce operation to each pair (key, list-of-values) in parallel. Figure 16 summarizes the application of the different MapReduce steps to our example.

Now, you must understand that the MapReduce approach is to reformulate its problem into parallelizable functions with a relatively constrained schema. But, what is its concrete implementation in a big data context? How does MapReduce respond to the main problems of distributed computing?

This framework supports distributed execution while ensuring: the access and sharing of data; the error handling and fault tolerance; and the location of the data.

Table 2. Map function application

Price of the first car:	$P1$	FunctionM (element1)
Price of the second car:	$P2$	FunctionM (element2)
Price pf the third car:	$P3$	FunctionM (element3)
Price of the fourth car:	$P4$	FunctionM (element4)

It is obviously necessary to understand and control the functioning of these distributed execution frameworks for the implementation of MapReduce algorithms, but finally, the first big difficulty is the reformulation of your problem in MapReduce. To practice thinking in MapReduce, I suggest you illustrate it again with this new example.

We consider the following four fictitious n-tuples (Table 1).

Calculate the maximum, average or total price can be written using algorithms, of the type:

```
for each n-tuples, calculate:
        value = FunctionM(current n-tuple)
return          FunctionR(values encountered)
```

For example, FunctionM extracts the price of the car, and FunctionR calculates the max of a set of values:

```
for each car, calculate:
        price = getPrice(current car)
return max(price met)
```

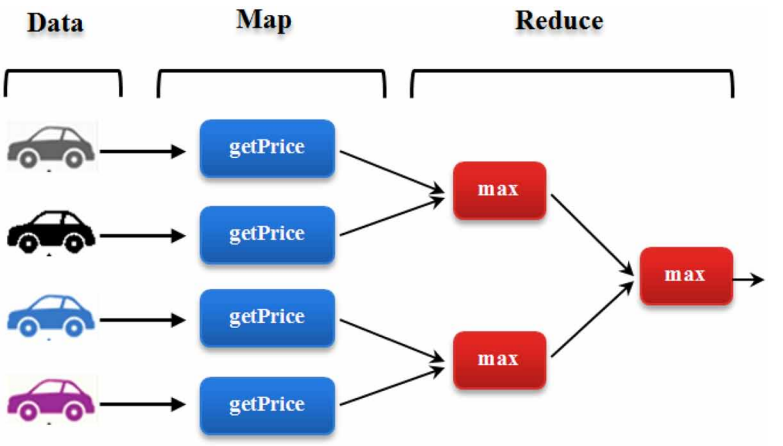
For efficiency, the intermediate values are not stored but transmitted between the two functions.

The map function is inherently parallelizable because the calculations are independent. For the four elements to treat, see Table 2.

Table 3. The Reduce phase

Max P1 & P 2 =	FunctionR (value1, value2)
Max P3 & P4 =	FunctionR (value3, value4)
Result =	FunctionR (Max P1 & P 2, Max P3 & P4)

Figure 17. MapReduce design results



The calculations can be done simultaneously, for example on four different machines. It should be noted here, that the map function is a pure function of its parameter, that it has no edge effect such as modifying a global variable or memorizing its previous values.

The reduce function is partially parallelized, in a hierarchical form, for example (Table 3).

Only the first two calculations can be done simultaneously. The third (the result) must wait. If there were more values, we would proceed as follows:

- Parallel computation of the FunctionR on all the pairs of values coming from the map
- Parallel computation of the FunctionR on all the pairs of intermediate values resulting from the preceding phase.
- And so on, until only one value remains.

CONCLUSION

These are just a few examples of applications among many others, as data analytics methods can be applied in many tasks requiring pattern recognition or prediction. The multiplication of data produced today is a great advantage for the development of data analytics process that needs to 'feed' from this large amount of data in order to derive models and generate value.

If the whole process has worked well, it must still normalize everything and make it a routine because big data must one day be part of the dashboard of all entrepreneurs. It is the Excel of tomorrow. An environment dedicated to prediction and value creation. Big data analytics is the weapon of tomorrow's business playground, a field with many data management tools and advanced techniques that were still missing and that must be mastered now. If software gave the possibility to make projections or modeling the past to make scenarios of the future, then big data will make anticipation, this is to say: model the future to make good decisions today.

REFERENCES

Brynjolfsson, E., & Mcaffee, A. (2012). Big data: The management revolution. *Harvard Business Review*, *90*(10), 60–68. PMID:23074865

Davenport, T. H., & Kim, J. (2013). *Keeping Up with the Quants: Your Guide to Understanding and Using Analytics*. Harvard Business Review Press.

Gibson, D., & Ifenthaler, D. (2017). Preparing the next generation of education researchers for big data in higher education. In B. Kei Daniel (Ed.), *Big data and learning analytics: Current theory and practice in higher education* (pp. 29–42). Berlin: Springer. doi:10.1007/978-3-319-06520-5_4

Gibson, D., & Webb, M. (2015). Data science in educational assessment. *Education and Information Technologies*, *24*(4), 697–713. doi:10.100710639-015-9411-7

Lane, J. E. (2014). *Building a Smarter University: Big Data, Innovation, and Analytics*. New York: SUNY Press.

Sedkaoui, S. (2018a). *Data analytics and big data*. London: ISTE. doi:10.1002/9781119528043

Sedkaoui, S. (2018c). How data analytics is changing entrepreneurial opportunities? *International Journal of Innovation Science*, *10*(2), 274–294. doi:10.1108/IJIS-09-2017-0092

Walwei, U. (2016). Digitalization and structural labor market problems: The case of Germany. *ILO Research Paper*, *17*, 1-31.

KEY TERMS AND DEFINITIONS

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.

Classification: In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

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.

Correlation: A mutual relationship or connection between two or more things.

Entrepreneur: Entrepreneurship is not only an outcome of the ecosystem but also an important input factor since entrepreneurs drive the ecosystem by creating it and keeping it healthy. Drucker believes that what entrepreneurs have in common is not personality traits but a commitment to innovation. For innovation, the entrepreneur must have not only talent, ingenuity, and knowledge but he must also be hardworking, focused, and purposeful.

MapReduce: 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.

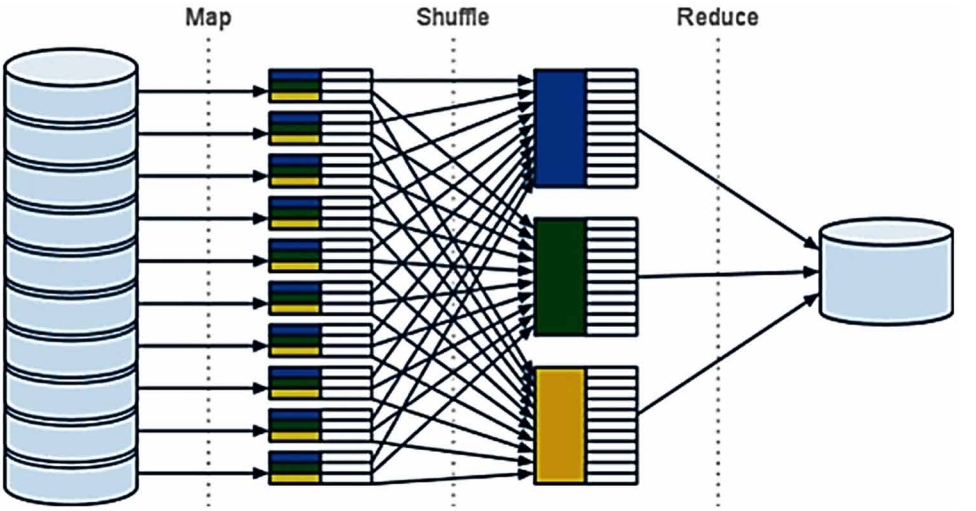
Missing Values: Occur when no data value is stored for the variable in an observation.

Outliers: An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

Regression Analysis: Is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables when the focus is on the relationship between a dependent variable and one or more independent variables (predictors).

APPENDIX

Figure 18. MapReduce model



Conclusion

There is a way between voice and presence, where information flows. In disciplined silence it opens; with wandering talk it closes. (Rumi)

We are currently living a great time, which allows a rapid diffusion to a large number of extremely powerful technological tools and analytics algorithms. Big data analytics echoes recent advances in this area, helping to diffuse them into the business sphere and helping to improve decision-making processes by streamlining them on solid foundations.

So, every entrepreneur is concerned about the arrival of these new powerful tools and algorithms. Analytics widens entrepreneurs or small businesses scope as an entity, giving them the ability of doing things they never thought were possible, for example, it offers timely insights, to allow them in making better decisions, about business innovation opportunities, it also helps them in asking the right questions and supports them with extracting the right answers as well.

Whether if they understand the potential of big data or not, or if they want to embrace the analytical IT tools or not, build a career in data science or not. One thing is confirmed and cannot be avoided: big data analytics can fundamentally change the way the activities operate. In this context, an entrepreneur can see new opportunities, manage and increase business value, by putting the efforts in the right direction, and rationally use his time and energy.

The data revolution continues... And eventually, we will all become interesting by being "an entrepreneur made in data". And, if the *Harvard Business Review* has called the data scientist as "*the sexy new job of the 21th century*"; then data-driven entrepreneur, who can understand data and have a strong creative streak in order to ask the right questions to get significant

value from data, will be the *fashionable* and the *chic* entrepreneur of the century. So, for entrepreneurs, or for those who want to undertake in the big data universe, for those who have a critical vision towards how generate value or the 'power' from data, prepare yourself to the data revolution age: the age where many underlying messages can be transformed into opportunities in order to help in addressing business challenges.

Here we have already crossed the end of this book. I hope that you have enjoyed reading it and that it have given you a general overview of the methods and applications of big data analytics.

During its reading, you have to understand that data analytics, if it feeds on the latest advanced technology, is accessible to anyone with a minimum of analytical and statistical background and some business keys to understanding the universe and how it works. We hope that this book has opened new horizons to you, by presenting new approaches that you may not have known before.

We also hope that it has helped to sharpen your curiosity and stimulate your desire to learn more about it and continue to your learning process. It is a new revolution and a new deal. It is up to you now to show your know-how! The analytical tools will help you but as the Americans say: "*People make the difference*!"

About the Author

Soraya Sedkaoui is a Senior Lecturer and Data Analyst with more than 10 years of Teaching, Training and Research experience in econometrics, statistics and big data analytics. She earned her PhD in Economic Analysis and HDR in Economic and Applied Statistics. She was working as a Researcher at TRIS laboratory, University of Montpellier, France (2011-2017). Her science-oriented research experience and interests are in the areas of big data, Computer Science and the development of algorithms and models for business applications and problems. Dr. Sedkaoui's prior books and research has been published in several refereed editions and journals.

Index

G

H

I

K

M

N

O

P

R

S

T

U

W

Y

Z