INNOVATION, ENTREPRENEURSHIP, MANAGEMENT SERIES

SMART INNOVATION SET

Volume 3

Big Data, Open Data and Data Development

Jean-Louis Monino
Soraya Sedkaoui

iSTE
WILEY

Big Data, Open Data and Data Development

Smart Innovation Set

coordinated by
Dimitri Uzunidis

Volume 3

Big Data, Open Data and Data Development

Jean-Louis Monino
Soraya Sedkaoui

ISTE

WILEY

First published 2016 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

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

www.wiley.com

Library of Congress Control Number: 2016931678

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN 978-1-84821-880-2

Contents

Acknowledgements

This book is the product of several years of research devoted to data processing, statistics and econometrics in the TRIS (*traitement et recherche de l'information et de la statistique*) laboratory. It is the fruit of several projects carried out within the framework of research and development for several startups within the Languedoc-Roussillon region and large private and public groups.

I would like to thank all of the members of the RRI (*réseau de recherche sur l'innovation*), and more particularly, Dimitri Uzunidis, its president, for his attentive and careful reading of the first version, and who encouraged us to publish this book.

Thanks also to M. Bernard Marques, who had the difficult task of proofreading the manuscript and who had many important notes to help with the understanding of the book.

I would also like to thank my teacher and friend Jean Matouk, who was the cause of this publication, thank you for his encouragement and unfailing support over the years.

Many thanks to all the researchers at the laboratory for their help and support and most especially to Soraya Sedkaoui; without her this book would never have seen the light of day.

Thanks to all those who have supported me through difficult times and who have transformed an individual intellectual adventure into a

collective one, in particular Alain Iozzino, director of the startup E-prospects, with whom we have carried out many research and development projects over the years.

Finally, I must express my special gratitude to those dear to me, my family, and most of all to my wife, who has had to put up with my moods over the last few years.

Jean-Louis MONINO

This book was a work of adaptation, updating and rewriting in order to adapt all of the work of the TRIS laboratory. Its creation was fed by exchanges and discussions with my teacher Jean-Louis Monino, without whom this book would never have seen the light of day. I am infinitely grateful to him for having included me in this adventure.

It would not have been possible to produce this book without my family who have always encouraged and supported me throughout all my ideas and projects, no matter how far away they have sometimes been. Special mention must go to my mother, there are no words to express how important she is and how much she has done in making me what I am today.

Finally, I would like to thank Hans-Werner Gottinger, Mohamed Kasmi and Mustapha Ghachem for their unfailing support and for the interest that they have always shown in what I am doing.

Soraya SEDKAOUI

Foreword

The world has become a digitalized place, and technological advancements have multiplied the ways of accessing, processing and disseminating data. Today, new technologies have reached a point of maturity. Data is available to everyone throughout the planet. In 2014, the number of Internet users in the world was 2.9 billion, which is 41% of the world population. The thirst for knowledge can be perceived in the drive to seize this wealth of data. There is a need to inquire, inform and develop data on a massive scale. The boom in networking technologies – including the advent of the Internet, social networks and cloud computing (digital factories) – has greatly increased the volume of data available. As individuals, we create, consume and use digital information: each second, more than 3.4 million emails are sent throughout the world. That is the equivalent of 107,000 billion emails per year, with over 14,600 per person per year, although more than 70% of them are junk mail. Millions of links are shared on social networks, such as Facebook, with over 2.46 million shares every minute. The average time spent on the Internet is over 4.8 hours per day on a computer and 2.1 hours on a cellphone. The new immaterial substance of "data" is produced in real-time. It arrives in a continuous stream flowing from a variety of generally heterogeneous sources. This shared pool of all kinds of data (audio, video, files, photos, etc.) is the site of new activities aimed at analyzing the enormous mass of information. It thus becomes necessary to adapt and develop new approaches, methods, forms of knowledge and ways of working, all of which involve new paradigms and stakes as a new

ordering system of knowledge must be created and put into place. For most companies, it is difficult to manage this massive amount of data. The greatest challenge is interpreting it. This is especially a challenge for those companies that have to use and implement this massive volume of data, since it requires a specific kind of infrastructure for the creation, storage, treatment, analysis and recovery of the same. The greatest challenge resides in "developing" the available data in terms of quality, diversity and access speed.

Alain IOZZINIO
E-PROSPECTS Manager
January 2016

Key Concepts

Before launching into the main text of this book, we have found it pertinent to recall the definitions of some key concepts. Needless to say, the following list is not exhaustive:

– *Big Data*: 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, speed, and variety are usually the three criteria used to qualify a database as "Big Data".

– *Cloud computing*: This term designates a set of processes that use computational and/or storage capacities from remote servers connected through a network, usually the Internet. This model allows access to the network on demand. Resources are shared and computational power is configured according to requirements.

– *Competitive intelligence*: It is the set of coordinated information gathering, processing and dissemination activities useful for economic actors. According to the Marte Report, competitive intelligence can be defined as the set of coordinated information research, processing and dissemination actions aimed at exploiting it for the purpose of economic actors. This diverse set of actions is carried out legally with all data protection guarantees necessary to preserve the company's assets, with the highest regard to quality, deadlines and cost. Useful information is needed at the company or partnership's different decision-making levels in order to design and put into place strategies and techniques coherently aimed at achieving company-

defined objectives and improving its position in the competitive environment in which it operates. These kind of actions take place in an uninterrupted cycle that generates a shared vision of company objectives.

– *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 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 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 governance*: It constitutes a framework of quality control for management and key information resource protection within a company. Its mission is to ensure that the data is managed in accordance with the company's values and convictions, to oversee its quality and to put mechanisms into place that monitor and maintain that quality. Data governance includes data management, oversight, quality evaluation, coherence, integrity and IT resource security within a company.

– *Data journalism*: The term designates a new form of journalism based on data analysis and (often) on its visual representation. The journalist uses databases as his or her sources and deduces knowledge, meaningful relationships or intuitions from them that would not be accessible through traditional research methods. Even when the article itself stands as the main component of the work, illustrating ideas through graphs, diagrams, maps, etc., is becoming more important day by day.

– *Data mining*: Also referred to as knowledge discovery from data, is intended for the extraction of knowledge from large amounts of data using 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 reuse*: This practice consists of taking a dataset in order to visualize it, merge it to other datasets, use it in an application, modify it, correct it, comment it, etc.

– *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.

– *Data visualization*: Also known as "data viz", it deals with data visualization technology, methods and tools. It can take the form of graphs, pie-charts, diagrams, mappers, timelines or even original graphic representations. Presenting data through illustrations makes it easier to read and understand.

– *data.gouv.fr*: The French government's official website for public data, which was launched on December 5th 2011 by Mission Etalab. In December 2013, data.gouv.fr was transformed deeply through a change in both the site's structure and its philosophy. It has, without doubt, become a collaborative platform oriented towards the community, which has resulted in better reuse of public data.

– *Dataset*: Structured and documented collection of data on which reusers rely.

– *Etalab*: This is a project proposed in the November 2010 Riester Report and put into place in 2011 which is responsible for implementing the French government's open data policy, as well as for establishing an almanac of French public data: data.gouv.fr.

– *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 describes and answers questions like “who?”, “what?”, “when?” and “how many?”.

– *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 involving significant changes in a product’s design, conditioning, placement, promotion or pricing.

- *Organizational innovation*: Establishing a new organizational method in practices, workplace organization or company public relations.

– *Interoperability*: This term designates the capacity of a product or system with well-known interfaces to function in sync with other existing or future products or systems, without access or execution restrictions.

– *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.

– *Linked Open Data (LOD)*: This term designates a web-approach proposed by supporters of the “Semantic Web”, which describes all data in a way such that computers can scan it, and which links to it by describing its relationships, or by making it easier for the data to be related. Open public data is arranged in a “Semantic Web” format,

such that its items have a unique identifier and datasets are linked together by those identifiers.

– *Open innovation*: It is defined as increased use of information and knowledge sources external to the company, as well as the multiplication of marketing channels for intangible assets with the purpose of accelerating innovation.

– *Open knowledge foundation network*: A British non-profit association that advocates for open data. It has most famously developed CKAN (open source data portal software), a powerful data management system that makes data accessible.

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

– *Semantic Web*: This term designates a set of technologies seeking to make all web resources available, understandable and usable by software programs and agents by using a metadata system. Machines will be able to process, link and combine a certain amount of data automatically. The semantic web is a set of standards developed and promoted by W3C in order to allow the representation and manipulation of knowledge by web tools (browsers, search engines, or dedicated agents). Among the most important, we can cite:

- *RDF*: a conceptual model that makes it possible to describe any dataset in the form of a graph in order to create knowledge bases;

- *RDF Schema*: language that makes it possible to create vocabularies, a set of terms used to describe things;

- *OWL*: A language that makes it possible to create ontologies and more complex vocabularies that serve as support for logical processing (interfaces, automatic classification, etc.);

- *SPARQL*: A query language for obtaining information from RDF graphs.

– *Semi-structured information*: It is worth noting that the boundary between structured information and unstructured information is rather fuzzy, and that it is not always easy to classify a given document into