ATLANTIS PRESS

Sander Klous
Nart Wielaard

We are Big Data

The Future of the Information Society

We are Big Data

Sander Klous · Nart Wielaard

We are Big Data

The Future of the Information Society

ATLANTIS PRESS

Sander Klous
Informatics Institute
University of Amsterdam
Amsterdam, North Holland
The Netherlands

and

Management Consulting
KPMG
Amstelveen, North Holland
The Netherlands

Nart Wielaard
Nart BV
Haarlem, North Holland
The Netherlands

ISBN 978-94-6239-182-6 ISBN 978-94-6239-183-3 (eBook)
DOI 10.2991/978-94-6239-183-3

Library of Congress Control Number: 2016937342

Printed on acid-free paper

 [1]

Art becomes art when you can see the complete picture.
Data acquires value when you can understand the context.

[1]What is this? It will become clear as you read this book.

Foreword

Big Data matters—a lot. But in a more subtle and fundamental way than is often portrayed today. Marketers from consulting and IT companies tout the virtues of Big Data with deafening volume, limited variety, and increasing velocity. Their mantra is that with new IT and a consulting partner lined up, any company can turn itself into a Big Data superpower. That is patently wrong—and the marketers' behavior is as predictable as it is irresponsible.

By the same token, commentators and self-styled cyber-experts paint dire monochromatic pictures of a Big Data future in which everything human will have become eliminated. Listening to some of them, it seems that with Big Data, we'll turn our computers into weapons of mass destruction. Such doomsday predictions may sell books but, just like the Big Data hype do little to improve our understanding of what Big Data actually is—and how fundamentally it will change our economy and our society.

Big Data is, at its very core, nothing more than process, a unique mechanism of how we humans make sense of the reality around us and, based on this understanding, make predictions about our future that are far more accurate than the tea-leaf reading of the past.

For all of human history, we have made sense of the world that surrounds us by watching it and thinking about it—by theorizing about how reality comes together. And we gathered data about the world to prove (or disprove) our theories. Over time, we realized the need to see the world in rational terms, and to engage

in understanding it in a methodical fashion. This is the seedling that bloomed in the Age of Enlightenment and which was nurtured among others by Spinoza, one of Holland's intellectual heroes. Eventually this yielded the "scientific method" and the great progress in understanding and utilizing the world in the 19th and 20th centuries. As a result, many more people have been living longer and better.

But until recently, the collection, analysis, and storage of data was time-consuming and costly. Thus, we collected as little data as possible to answer the questions we had. In fact, in the age of small data, our institutions and processes the very way we gained insights from data was premised on data scarcity. So we used samples of data rather than all of it—for everything from quality control in manufacturing to polling voters before elections. That shortcut worked if done correctly (which was not always the case), but it lacked details. So we could rarely look deeper into something that turned out to be interesting, because we simply did not have enough data.

More troubling, with such an approach we could only answer questions we already had at the outset, not use the data effectively to pose new questions that might lead to truly novel insights. We also privileged data quality, striving even at high cost to improve the accuracy of the little data we collected.

Today, many of the barriers to collecting, analyzing, and storing large amounts of data have been greatly reduced. This makes it vastly easier to collect all, or close to all, the data and capture the complexity of our world rather than settle for just a small sample of reality. In many cases we can also choose to collect more, but messier, data rather than gather very small amounts.

As a result, we can now look at the world in unprecedented detail and Big Data analysis yields new questions that point us toward valuable insights we otherwise would not have. This is how Google is attempting to predict the spread of the flu by analyzing the billions of search requests it receives. This is how Rolls Royce can foresee when a part in one of its airplane engines is likely to break before it actually does, which not only saves money but also saves lives.

And this is how startup companies like INRIX are able to direct their customers around traffic jams on their daily commutes as well as advise local governments on emerging traffic patterns. This makes IBM's Watson computer win against the best humans in *Jeopardy*, and help doctors in their diagnoses. At the University of Ontario they have even found a way to save the lives of premature babies harnessing Big Data. In the Big Data age, more and messy data has taken over from clean, but small, data.

But there is a third important shift in how we make sense of the world. We humans prefer to understand the world as a sequence of causes and effects. And so we cannot help but look for causes of everything we notice—unexpected noises, a car that won't start, a shopper's decision to buy, the way we feel. But as it turns out, identifying causes is hard. So, as Nobel laureate Daniel Kahnemann has shown, we often tend to intuit causes without conclusive evidence. That comforts us and it gives us a false sense that we understand the world.

Much of Big Data analysis cannot tease out causes from analyzing effects. It regularly cannot tell us the "Why", only "What". But, amazingly, that is often good enough to improve our decision-making. It helps Google translate dozens of languages on the fly and accounts for about a third of Amazon's sales. It even lets cars drive themselves.

This of course does not mean that our lifelong quest for causes suddenly should come to an end. Rather, it helps us realize that before we can run, we need to be able to walk. With Big Data we have a powerful tool to do just that. If applied appropriately, Big Data analysis will discover the "What", thus acting as an invaluable filter for subsequent thorough causal analyses.

This helps us appreciate that the world we live in is far more complex, but also far more interesting and thought-provoking than we thought. As we realize that the data we can now gather holds so much latent value which we can uncover through Big Data, businesses will rethink how they operate, reshaping our economy and helping our society evolve.

In this smart and insightful book, written by experts who truly understand and grasp Big Data, you will gain a new appreciation for Big Data and how we can harness it for the good. It offers a passage to an exciting journey that can yield amazing insights and produce significant value. Importantly, you will also hear how Big Data's power can be channeled so that we can benefit from it and avoid suffering from some of its negative consequences.

Big Data matters—a lot. Make it matter for you!

Viktor Mayer-Schönberger
Co-author of Big Data: A Revolution That Will
Transform How We Live, Work, and Think

Contents

Foreword vii

Introduction: Anything is Possible xiii

1 Big, Bigger, Biggest Data 1

2 Remove Fear and Conquer Resistance 17

3 You Ain't Seen Nothing Yet 43

4 Hitting the Bullseye First Time Around 53

5 Data Stimuli for a Better World 77

6 Data Analytics Is the Society 95

7 Wanted: Thousands of Sherlock Holmes Clones 107

8 The Question Is More Important Than the Answer 115

9 Transparency, a Weapon for Citizens 131

10 Systems Determine Our Behavior 145

11 A New Ecosystem 157

Epilogue 173

Word of Thanks 179

Notes 181

Index 193

Introduction: Anything is Possible

Expecting the extreme is the new norm

The iPad Generation

YouTube[1]. We see a baby—so young that she can only coo—in front of several glossy magazines. She is turning pages and wants to zoom in or click through. The usual swiping, pinching and tapping— she has obviously already mastered on a tablet at this young age—isn't working. Disappointed, she throws the magazines aside. This was filmed by her father, Jean-Louis Constanza, then-CEO of Orange Vallée, who sums it up beautifully: "In the eyes of my one-year-old daughter, a magazine is a broken iPad. That will remain the case throughout her entire life. Steve Jobs programmed part of her operating system."

New technology has made many incredible things possible. It is now part of our lives in a way we couldn't even have dreamed of 20 years ago. Okay, no more clichés…

Don't worry; this book isn't (really) about technology.

We could throw around some impressive numbers about the phenomenal calculative capacity required for the successful search for the Higgs boson particle—the 'God particle'—at CERN, the European organization in Geneva that conducts fundamental research into elementary particles and is therefore in the Champions League of data analysis. (One of the authors, Sander, was involved with the Atlas experimentat CERN, which analyzed the data obtained by the particle accelerator.) We could explain that new

open-source products are the basis of numerous new groundbreaking data applications. We could attempt to unravel the algorithms that allow Google to return such good results for your searches.

The New Information Society

Each of these is an interesting technological subject, but we only discuss them briefly in this book. We prefer to dwell on the impact of this kind of technology on society and, more specifically, how this technology, related to data, data processing, information, and communications, can improve our society. A new world is rapidly emerging, a new world that, in this book, we call the new information society. It is a world in which everything is measurable and in which people, and almost every device you can think of, are connected 24/7 through the Internet. That network of connections and sensors generates a phenomenal amount of data and offers fascinating new possibilities that, together, are often called Big Data. Without a doubt, these possibilities are vast and they will be exploited. They're already being exploited! We are in the middle of this change, even if we are not always aware of it.

Awareness of this change is spreading through businesses and governments who are asking themselves how to respond to these developments. This goes beyond launching new products or services. It's about how to respond to a totally different world. In this respect, the term 'age of disruption' is increasingly being used. The term describes the disruptive effect that rapid and successive technological breakthroughs are having on virtually all aspects of society, which is definitely challenging, but also offers enormous opportunities.

In macroeconomic terms, there is reason for great optimism. We are in the middle of a classic economic phenomenon: 'the creative destruction'.[2] This means that we can only achieve innovation once the old order has been dismantled, which is already happening in almost all sectors. Big Data is acting as a growth hormone for this creative destruction. The growth is undeniable. All our chatter on social media, online music and video streaming, our e-mails, online shopping transactions and so on result in barely comprehensible

amounts of data traffic. And that mountain of data is growing rapidly now that the 'Internet of Things' is linking billions of devices which were previously 'offline'—TVs, refrigerators, security devices, thermostats, smoke detectors—all of which now produce and share data.

It's tempting to start throwing impressive figures around here, but that's not the purpose of this book. One figure is actually enough: in the last two years, the amount of data recorded was 10 times greater than the amount previously recorded over the entire history of humanity.[3] This seems impressive, but what's really impressive is what you can do with that data. That is what we discuss in this book.

The amount of data produced is growing exponentially, to astronomical levels, yet the word 'big' in Big Data is actually slightly misleading. Many new applications are not about editing or interpreting enormous amounts of data—the so-called 'big, messy data'—but much more about the smart combination of data.

An example is the rise of the 'Quantified Self' phenomenon,[4] the trend where an increasing number of people are measuring numerous aspects of their lives. They want to gain insights into what they're eating, how much they're moving, how deeply they're sleeping, what happens to their heart rates when they exercise. They themselves determine the level of detail they want to measure. For example, they can record how much coffee they drink or how much chocolate they eat. Through various apps on smartphones or other devices, they can simply record information on their own behavior, health and lifestyle and can then use this data to motivate themselves, for example, to reach the next level of fitness. Semantically, this has little to do with *Big* Data. However, it forms part of what society considers Big Data.

Big Data in This Book

We have chosen to use Big Data as an umbrella term. Big Data, in this book, includes all the new opportunities, possibilities, techniques, and threats associated with the fact that we can now deal with data

differently or, in other words, the positive and negative aspects of the 'datafication' of society, including social facets such as privacy and social influences. This includes the extensive monitoring of personal behavior by intelligence agencies, commercial precision-bombing of customers, resolving traffic congestion, stopping epidemics and connecting a refrigerator or thermostat to the Internet. An important part of Big Data is data analysis: searching for patterns in the large amounts of data to enable relevant conclusions to be drawn.

Relationship with Technology

We believe that Big Data can be extremely valuable to the world. We are already experiencing some of the amazing benefits every day, as technology has touched almost every aspect of our lives. We are all familiar with the technology used to communicate with friends and family all over the world. We can send them text messages on smartphones, play online games with them on our tablets or keep in touch via social media. We can also take care of practicalities through the wonders of technology, like submitting bills from our physiotherapists straight to our health insurers over the Internet.

We store photos and videos in the cloud—which is very handy when we want to share them—and we use the cloud to avoid holding all of our business data on our own hard drives. Keeping things online is easier and reduces the risk of us losing them. We arrange our banking with an app. When we need information about our cocker spaniel's sudden symptoms, we find it on the Internet in an instant. When we're arguing in a bar about whether Peter O'Toole did or didn't win an Oscar, it takes 10 seconds to find the answer with a smartphone. As technology entrepeneur Peter Hinssen said in one of his books: "Digital is the new normal in everything we do. Digital living is not only something we do completely routinely, we have also become addicted to it and our expectations about the possibilities of this digital life have increased tremendously."[5]

Expectations

Where will it end? What will digital life look like in the future? Are we all going to be wearing Google Glass—or its decendants—as a replacement for our smartphones? Or will the smart watch break through? Perhaps. Will employment decrease by half over the next 50 years because more and more tasks, such as writing a book like this, will be taken over by computers? Maybe. Will we be able to extend our average life-span by 20 years, in part because we have the computing power to map the human genome? It's possible. But the honest answer is that nobody can say with any certainty how the use of technology will evolve.

We have no doubts that technology—and hence Big Data—will infiltrate every aspect of our lives even more in the years to come. Not only because it's impossible to avoid, but also because we actively want it. Because we don't want to live without the convenience, comfort, or added value that technology brings.

Technology creates expectations, like the toddler who expected to use a paper magazine as a tablet. We should be able to do anything; have an app for everything. The uptime of all services needs to be 100 %. Companies will need to provide tailor-made products, preferably at zero cost. Moreover, it will all need to be fun and sustainable.

All of this makes it easy to summarize the new challenge facing companies, governments, and other organizations: *the impossible needs to be available immediately; delivery within two weeks is permissible for miracles.*

Possibilities

Big Data is more than just a new technology or concept. It offers opportunities to all sectors and industries for us to organize ourselves differently, to make progress and to do things that were simply not possible until recently. Many people associate the term 'Big Data'

with companies wanting to sell their customers even more stuff by learning everything there is to know about them. However, that's only one side of the coin, a side that we will definitely address in this book. We also want to show that Big Data can solve societal problems and improve our lives. We can use it to save human lives. We can improve maintenance and plan more cost-effectively. We can improve traffic safety and reduce congestion. We can increase agricultural yields. We can provide a better quality of life for the elderly. We can make the world more sustainable. We can revolutionize medical diagnostics and treatments. We can increase the security of payments. We can even improve a team's sporting performance.

Structure

We begin this book with an overview of what we can do with Big Data and elaborate on both the benefits and the drawbacks. We discuss how we, as citizens and consumers, have extreme expectations, but we also feel resistance against the appetite of organizations for our personal data and we fear for our privacy. In doing so, we reach the conclusion that we cannot deal with Big Data in isolation, because it is fundamentally linked to social progress.

That's why in the second part of the book we discuss the new information society that is emerging, of which Big Data is an essential part. We show that individuals are increasingly demanding that organizations give back to society rather than acting only for gain, that we will become ever more powerful as a collective of individuals and that, as a collective, we can have a significant influence on the actions of organizations. In this new world, we must find a way to embed Big Data successfully into our society.

Broadening our scope to take in the information society as a whole leads, in the third part of the book, to a discussion of how Big Data can be successfully embedded into society. Issues that are addressed include ethical principles, education and the role of government.

Big Data can—in a controlled manner—contribute to a better society. That's exactly why we wanted to write this book. We want to show that companies and governments still don't know how to deal with Big Data. The subject is dogged by misunderstanding and ignorance. Big Data projects are often fragmented and unstructured. This leads to unsuccessful projects, significant disappointment, and resistance. This must change. Throughout this book, we aim to show how things could be different. In the following chapter, we first look at the (potential) advantages that Big Data has to offer us.

Chapter 1
Big, Bigger, Biggest Data

New opportunities, made possible by an abundance of data

Will It Ever Be Possible to Predict Crime?

Actor Tom Cruise storms into a house in the suburbs of Washington and runs straight to the bedroom, just before the resident tries to stab his wife—whom he found in bed with another man—in the chest. Cruise states his name is John Anderton, chief of a police unit focused on crime prevention, and arrests the man for the murder he was about to commit. This opening scene of the movie Minority Report suggests that in the year 2054, predictions will be so exact that people will be arrested before they commit a crime.

Will it ever be possible to predict crimes through data analysis? Should we want to? The reality is that several American cities are already successfully using 'predictive surveillance'. Based on the analysis of large amounts of data from various sources, algorithms determine which streets require additional attention from police officers. According to Predpol, one of the key suppliers of predictive policing software, the effect will be more arrests, but also lower crime rates. An experiment conducted in a town in Kent, England showed that number crunching is significantly better at predicting where a crime will occur than the conventional approach. Data analysis made a correct prediction in 8.5 percent of cases, whereas predictions made by experienced professionals scored no better than 5 percent. In an earlier experiment in Los

S. Klous and N. Wielaard, *We are Big Data*,
DOI 10.2991/978-94-6239-183-3_1

Angeles, California, these scores were 6 percent and 3 percent. *The Economist*, a British weekly magazine, had a pointed title for an article about this: "Don't even think about it".[1] However, some experts have expressed criticism regarding the expediency and possible side effects of such measures. Namely, an individual's right to privacy could lose out to society's need for safety.[2]

In another example, again from the US: Baltimore's Parole Board uses risk profiles that indicate the level of risk of a detainee becoming the perpetrator or victim of a murder after being released from custody.[3] This risk calculation forms the basis for deciding what level of monitoring is required for the detainee about to be released.

These are all examples that show how we can increase societal safety by analyzing data in a smart manner. Restraint is sometimes required and policy transparency is a must. For example, in the case of releasing detainees based on data analyses, it is important that the authorities clearly show the public how decisions are made.

Not New

Of course, Big Data is not completely new. For decades, research institutions and companies have gathered large amounts of data in order to generate new information. Stock exchange traders use models that draw on extensive data streams from various sources for the early detection of trading risks and opportunities. Intelligence agencies search for patterns in data in order to prevent potential attacks. Insurers use data to estimate the risk levels of individual customers and entire portfolios. The unifying theme in this respect has been the same for decades: we search for correlations, early indicators and cause-and-effect relationships between phenomena, persons and events and make decisions based on the results.

However, since the advent of the Internet, things have changed. The application of data analysis has broadened and deepened and has entered our daily lives. This is further accelerated through the rise of the 'Internet of Things' in which all kinds of appliances, big and small, are connected both with each other and with us.

How and why (the use of) data analysis is changing is discussed below.

Exponential Growth of Information Technology Possibilities

Gordon Moore, one of the founders of Intel, predicted in 1965 that the number of components in a dense integrated circuit would double every two years due to technological progress (Moore's law). In simple terms, every two years, twice the computational power could be bought for the same amount of money or the same computational power for half the amount. Although experts have been expecting for years that, as the limits of fundamental physics (such as the number of atoms per circuit) are reached, the prediction would cease to be true,[4] it still applies.

Comparable laws apply to network and storage capacity. In this respect, it is interesting to note that storage capacity doubles more quickly—within 12 months—than computational power—within 18 months. Since, generally speaking, we actually use that storage capacity, the conclusion is that increasingly less computational power is available per byte of stored data. In other words, data is becoming exponentially unanalyzable[5] with traditional methods, forcing a rethink on how we exploit the vast amounts of data we are producing.

As well as the exponentially increasing technical possibilities, another interesting shift is occurring. Until recently, research and progress in data analysis were mainly achieved by scientific

institutions; however, nowadays the big breakthroughs are increasingly being achieved as a result of major advances by the business world. Companies such as Google, Facebook, Yahoo!, Apple, and Amazon are playing an important role in this respect. This is on top of the technological breakthroughs that were already the domain of commercial entities such as Intel, IBM and Cisco, and it is driving an accelerated learning and development process. A clear sign of this cross-pollination is the announcement by CERN[6] that it is to use Google technology, the Google toolchain, to manage the data centers responsible for the analysis of the impressive amounts of data generated by the Swiss particle accelerator.

Although the term 'Big Data' suggests that an increase in storage capacity and computational force is the key motivating factor, in reality it is the increasing capacity of networks that is driving the revolution. Twenty years ago, the possibilities in respect of computer systems were limited: there were limits to the amounts of data you could transfer and the way you could transfer data between systems. Now, an almost limitless number of sources can be reached with very little effort. Applications on devices like smart phones—such as those for streaming music—are possible because of this development.

Don't Decide Based on What We Say, but on What We Do

Traditional market research is slowly disappearing due to Big Data in light of the new ways organizations have of getting much closer to groups of customers or stakeholders. Companies no longer have to ask what people think or perceive (i.e. through market research) and just hope they will tell the truth. Instead, they can just track us. Where do we go? What do we buy? What music do we listen to? What movies do we watch? Whom do we call? Now