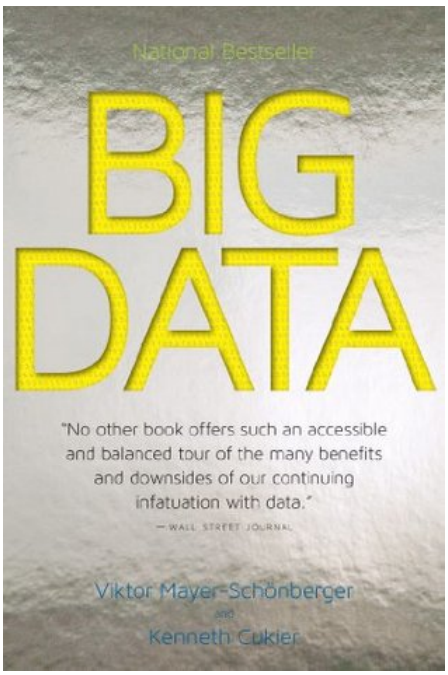

Big Data: A Revolution That Will Transform How We Live, Work, and Think

Viktor Mayer-Schönberger

Big Data: A Revolution That Will Transform How We Live, Work, and Think

Viktor Mayer-Schönberger

Big Data: A Revolution That Will Transform How We Live, Work, and Think Viktor Mayer-Schönberger
A revelatory exploration of the hottest trend in technology and the dramatic impact it will have on the economy, science, and society at large.

Which paint color is most likely to tell you that a used car is in good shape? How can officials identify the most dangerous New York City manholes before they explode? And how did Google searches predict the spread of the H1N1 flu outbreak?

The key to answering these questions, and many more, is big data. “Big data” refers to our burgeoning ability to crunch vast collections of information, analyze it instantly, and draw sometimes profoundly surprising conclusions from it. This emerging science can translate myriad phenomena—from the price of airline tickets to the text of millions of books—into searchable form, and uses our increasing computing power to unearth epiphanies that we never could have seen before. A revolution on par with the Internet or perhaps even the printing press, big data will change the way we think about business, health, politics, education, and innovation in the years to come. It also poses fresh threats, from the inevitable end of privacy as we know it to the prospect of being penalized for things we haven’t even done yet, based on big data’s ability to predict our future behavior.

In this brilliantly clear, often surprising work, two leading experts explain what big data is, how it will change our lives, and what we can do to protect ourselves from its hazards. Big Data is the first big book about the next big thing.

Big Data: A Revolution That Will Transform How We Live, Work, and Think Details

Date	:	Published March 5th 2013 by Eamon Dolan/Houghton Mifflin Harcourt
ISBN	:	
Author	:	Viktor Mayer-Schönberger
Format	:	Kindle Edition 257 pages
Genre	:	Nonfiction, Business, Science, Technology

From Reader Review Big Data: A Revolution That Will Transform How We Live, Work, and Think for online ebook

Long Tùng says

Actually nothing is new in this book. Way too long and repetitive. The book should have been only 30 pages instead of 300.
Sorry but it's a waste of time.

Marks54 says

I chose to read this book because it is one of the more visible examples of a trade business/ technology book about some recent changes in the data/information business. The key intuition that this book is highlighting is a shift towards greatly increased production of data and greatly increased use of large nearly complete population levels data sets in the management and control of a range of industries. This change is fueled by the wide adoption of broadband internet services and significant increases in data processing capacities in pcs, laptops, and tablets. The key point of distinction is that when you can access nearly all of the data in a situation in real time, analysis approaches and techniques change and the uses of data analysis greatly increase. This moves situations away from traditional sample-based statistics and into much greater depth analyses and descriptions of entire populations (or as the authors say: n=all). The authors are from Oxford and "The Economist" magazine and appear to be experts, even though they claim in the book that experts will decline in importance in the big data sector.

The book is not without interesting sections, in particular the sections on Google's use of data for multiple purposes and the discussion of the value chain for the "big data" industry. This really is a change that is being discussed and business and engineering schools across the US are hurrying to catch up with and capitalize on it.

Having said that .... this book was a huge disappointment. Why?

To start with, the clear development that is the focus of the book about "big data" is blown way out of proportion and elevated to a near world historical event. I don't buy it for a second. The rise of data driven approaches to research and control is worth taking note of but claims similar to those made here have been made virtually nonstop for technological gains from Bell Labs through the Internet boom of the 1990s up to the present. It is as if each new fad needs its own slogan and fight song. Go big data, rah! To put it more simply, the hype levels are way over the top.

Second, even if I could put up with the hype, the book is not that informative beyond a very general level. Little depth is provided about firms like Google or Facebook. The most detailed examples are culled from other pop books with similar themes, including Freakonomics and Thinking Slow/Thinking Fast. A little more original contribution would have been appreciated.

Third, the book is rambling, repetitive, and poorly edited. I get it -- big data is awesome and will change everything! We have to work with correlations and not old fashioned ideas of causality! Once we reach n=all, our petty disputes will end and the facts will speak for themselves. The first four times these points were raised were sufficient for me.

Pay no attention to the fact that the bulk of examples provided are well understood examples of large technology firms, data intensive supply chain dominant firms (Wal-Mart), and internet start-ups. Pay no attention to the fact that many of these critical issues have been around for a while in various forms. Pay no attention to a historical record that shows that programming and data will not solve our problems, that technology company executives do not have all the answers, and that the facts do not speak for themselves.

The final chapters raise the obligatory issues of privacy, civil liberties, and controls, but it is hard to take such discussion seriously in a book that spent the initial chapters discussing what a cool development big data was.

The subject matter of the book is important and the changes discussed will contribute to changing our society. OK, but why not at least try to put things in a broader context? I really cannot recommend the book, but for interested readers there is an abriged version of the basic points in the May-June issue of Foreign Affairs.

Jimboninja says

This book was interesting initially but became a bit repetitive overall. It has three major points in it;

1. Sampling was important when collecting data was expensive and difficult, but we now we have access by one means or another to all data.

2. Since we have so much data, the quality of individual data points is not important and we can allow inexactness in measurement processes as long as there isn't a systematic bias.

3. Causality and understanding why things happen is no longer as important as correlation and discovering the correct strategies or course of action based upon large bodies of data.

All of these points could have been made in a hundred pages I think, and reading just the first half of the book would give a reader the basic ideas intended by the author.

While there may be areas in which data analysis can point out solutions to problems, I'm not convinced by the author's assertion that Big Data will make experts in various fields obsolete. I would guess that Big Data will be a tool in the hands of experts, but I don't think we'll find data analysts replacing doctors and subject matter experts on a large scale.

David Adams says

I just stopped reading this book at page 143 and I don't plan on continuing. The book is padded and repetitive plus most of the examples have already been published in articles. As another reviewer stated book... "it could've been reduced to a single feature-length article with links to prior stories"

Strange Statements:

-"Data has become a raw material for business... The data can reveal secrets to those with the humility, willingness, and the tools to listen"
Is this book about Data Science or did I just pick up Lord of the Rings?

vague definition of Big Data ie THE ISSUE of Big Data
"One way to think about the issue today - and the way we do in the book - is this : big data refers to things - one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value , in ways that change markets, organisations, the relationship between citizens and governments, and more..... this will change the way we live and interact with the world... society will need to shed some of its obsession for causality in exchange for simple correlations: not knowing why but only what."

???Is there anything solid there? Or is this a book about theory?

-BIG DATA overturns centuries of established practices and challenges our basic understanding of how to make decisions and comprehend reality.

-"At the beginning of the book he predicts that Silicon Valley's notorious hype cycle will kill off Big Data.... and the same people he states are so important to big data will flounder. Then vaguely aligns Big Data with invention of the telescope."

-Then .... ,"in this book we are not so much about big data's evangelists, but merely its messengers"

-"Big Data... Facts come in one end of the digital assembly line and processed information comes out at the other end-data this is starting to look like a new resource or factor of production"
(it call also resemble the digestive process as well)

- Defining Business Value for Big Data: "in the big data age, data is like a magical diamond mine that keeps giving long after its principal value has been tapped. (He later makes a argument counter to this.)

- He creates a Big Data Value Chain including an MVP role in Big Data... ie "The innovators and entrepreneurs"
-Then states: "the Big Data Gold mine is there for the taking .. " Yes direct reference to Big Data being a gold mine"

-Then of course identifies Internet Marketing as a niche / lucrative position within Big Data from an entrepreneurial perspective. Books like these are the reason we have 8 Billion Social Media Rockstar Gurus ....and also why uninformed people blow their hard earned life savings attempting to build the big data gold mine service 3.0

I've now made it 3/4 through this book with no reference to cloud, knowledge management, information architecture, semantic... nothing.

LOL.... I guess I expected more data less hype.

I didn't proofread this as I've wasted far too much time reading this book.

Rajat TWIT says

Thanks to the burgeoning Internet and the interwoven lives of people with the net world, we have so much of Data available that Google helped in a health emergency recently. That is how this book starts and this is the story it tells. Our every action of regular life creates a sort of data, which ultimately can be used to regulate, observe, predict and prepare a lot of things. From out air journeys to the bank transactions to the online searches, everything can be collected in the form of data, and can be used.
The crux of this book is an important learning and the authors have put in a lot of research work to enlighten the readers. The book is informative as well as thought provoking, it may even give you an idea or two to start a new venture of your own. Though the book is quite redundant and almost every second page has the same message: Big data will change the world. It gets annoying some times when even words and sentences are bearing the same essence. The authors could have kept the length of book to 150 pages, making it an interesting and not a redundant read.
A must read for those who handle data on regular basis, others may skip it after reading first two and last three chapters. They are most informative and important, though you will still find many redundancy in these chapters too, just like the word redundant getting redundant in this review. But I am doing this just to give you an idea how annoying this redundancy can be!

Liz says

Well-researched and well-written, this guidebook to the (dystopian?) future is intriguing, mind-bending, and alarming, all at the same time.

The authors explain that our economy and society are undergoing a massive upheaval, akin to the industrial revolution. “Representative sampling” used to be the norm, and researchers tightly controlled variables to isolate **why** something was occurring. Thanks to the massive increase in computer processing power, very inexpensive storage, and diverse types of data already in digital form (internet search terms, GPS coordinates, Facebook’s “social graph” of relationships, among others), the focus has now shifted to “bottom-up” analysis: i.e. looking at datasets (N= all) to detect patterns and insights, and see **what** is happening.

I kept coming back to 3 major concerns. First, we human beings are born asking “why?” If we substitute correlation (both A and B are occurring) for causation (A causes B), we end up with shortcut assumptions and abuses of existing data. (This is already happening, when employers use credit scores to evaluate job candidates, or health insurers review an applicant’s retail purchase history to determine underwriting risk.)

Second, our privacy is becoming not only diminished, but compromised in ways we can’t even foresee. Because data never wears out, it can be reused and re-examined in multiple ways (not to mention being repackaged and resold). As the authors say, “the three core strategies long used to ensure privacy – individual notice and consent, opting out, and anonymization – have lost much of their effectiveness.” On this last point, they cite an incident when AOL publicly released a dataset of 20 million search queries from 657,000 different users conducted between 3/1/06 and 5/31/06. Within days, reporters from the New York Times re-assembled searches to successfully id a particular user (a 62-year-old widow from Lilburn, Georgia).

Even mobile devices, which don't use "cookies," leave personal digital footprints. Tech companies (such as Drawbridge and Flurry) have sprung up expressly to track individuals, and can determine which devices (cellphone, work computer, home computer, tablet, laptop) belong to the same user.

The third concern (and here's where the dystopia really kicks in): authors Mayer-Schonberger and Cukier herald "the demise of the expert." They say: "many aspects of our world will be augmented or replaced by computer systems that today are the sole purview of human judgment." In Chapter 4, they introduce Dr. Carolyn McGregor, a Ph.D. computer scientist in Ontario, Canada who is creating software that analyzes vital sign monitor data to help doctors "make better diagnostic decisions when caring for premature babies." By Chapter 7, the authors are putting Dr. McGregor on equal footing with those experienced medical specialists: "[In a "small-data"] world, experience plays a critical role, since it is the long accumulation of latent knowledge – knowledge that one can't transmit easily or learn from a book, or perhaps even be consciously aware of – that enables one to make smarter decisions. But when you are stuffed silly with data, you can tap that instead, and to greater effect." Welcome to the "Moneyball" era of medicine.

As with any new methodology, the consequences of Big Data can be positive or negative, but are often unintended. In 2008, Barrack Obama's campaign manager used Big Data (from website visits) to mobilize grass-roots supporters and get his candidate elected. Two years later, republican-controlled state legislatures used Big Data (from the 2010 census) to redraw U.S. Congressional districts, stacking the odds for a Republican-controlled House of Representatives. In 2013, we end up with complete gridlock and a government shutdown.

Big Data = big opportunties, but big problems too. Use with caution.

Elizabeth Theiss says

Big Data takes on big ideas: how will the availability of huge data sets change the way we do research? Does it matter why something happens as much as what precisely is happening? Will correlation analysis of large data sets take the place of the painstaking theory-to-hypothesis-to-empirical-testing of traditional science?

The author perhaps over promises the impact of Big Data and underestimates the associated privacy issues. Yet, there is much that is interesting and important in this book. When N=All, a new set of issues arises and this book explores them substantively.

Big Data argues against many of Nate Silver's claims in The Signal and the Noise. Theory matters a good deal. Big Data points authoritatively and says Look Here. But it can't tell us what "it" means or why it matters.

I'm not sold on all the claims made by this book. The section on accountability seems weak and misguided. But I'm glad I read it and would recommend it to anyone who is curious about the impact of Big Data on scientific methods. Written for the lay audience, the book is engaging and clear.

Juliana says

This book has been sitting in my Kindle queue since publishing in 2013. Still holds up--but not a lot beyond talking at the high-level what can and has been done with big data. I didn't learn anything new.

Pradeep Nair says

The book tells you a lot about Big Data, the ubiquitous term nowadays.

There is so much of data being generated in the form of text, photos, and videos. Add to that the tons of personal data relating to whenever we do on our phone, like our location, what and how we are reading, listening, surfing the net, using different apps, etc.

Every minute detail of the way we use different apps are relayed back to the developers to get an understanding of the efficiency of the product. Everyone, not just Google, Amazon and the government are gathering data and analysing them, but everyone, including Goodreads.

Big Data is replacing the old cause-and-effect theory of 'if something is done in a particular way, it will have a particular effect", with correlation theory of "if most people are doing a particular thing in a particular way, then most others, if not everyone, are also likely to do that in the same way."

In the earlier era of small data, there was lot of importance to accuracy, but today, in the era of big data, there are more chances of inaccuracies, but that is compensated or nullified by the huge wealth of information that Big Data analyses provide.

The authors also rightly talk about the tyranny of data. Everything doesn't work according to numbers. There are many non-quantifiable and intangible, qualitative and contextual variables that affect analyses.

A good book, written in an easily understandable manner, especially for anyone who wants to know what Big Data is all about and how it's changing our lives.

On the flip side, the authors, in their attempt to explain different aspects of Big Data, tend to get too repetitive.

Trevor says

Chapter two of this book starts by telling us the three things the book will be about:

“The first is the ability to analyse vast amounts of data about a topic rather than be forced to settle for smaller sets. The second is a willingness to embrace data’s real-world messiness rather than privilege exactitude. The third is the growing respect for correlations rather than a continuing quest for elusive causality.”

And that is exactly what this book does – it discusses each of these points more or less in turn.

I hadn’t really thought about the relationship between statistics and big data before reading this book. Statistics is essentially a way of learning things about large populations by giving you an insight into how confidently you can make extrapolations from small samples of that larger population. So, you can ask a

thousand odd people how they are likely to vote in the coming election, and if you have randomised the people you are asking properly you are likely to get a pretty good idea of how the entire country will vote. This is a remarkable thing, when you think about it, and something very few of us (I include myself in this group) have ever really understood. Also remarkable is that asking another two or three thousand people the same question doesn't really make your prediction any more accurate, something else that is counter-intuitive with statistics that never made them feel all that comfortable. The word 'statistics' starts with the word 'state' for a very good reason – statistics have always been about finding ways to define and understand populations so as to control those populations.

But big data doesn't work in the same way that statistics generally works – big data isn't about asking a small group of people what their preferences are, big data asks everyone in the sample population what their preference is, in fact, big data doesn't even bother to ask, it just sees what we all do and goes from there. Rather than having to listen to our self-serving lies about ourselves, big data crunches how many times we actually eat at McDonalds, how many times we actually go to the gym, and draws conclusions from those facts, rather than our own wishful thinking.

So, we have supermarkets now that are able to link you (via your using their loyalty card) to everything you purchase in their store. You might also have told them other interesting bits of information: how much you earn, your address, your age, your marriage status, how many children you have, whether you own or rent your house – but then, I guess you were promised a Mars Bar at the end of each year, so why wouldn't you pass on this information? With the ever-increasing improvement in computer processing power this means that supermarkets are able to learn things about you that you probably don't know about yourself. All books on this subject (and this book is no exception) tell the Target story – where a father opens his daughter's junk mail from Target to find they are suggesting she buy various baby products. The father gets upset and calls Target to complain, only to later learn that his daughter actually is pregnant. How did Target know? They ran the girl's purchases through their computer and they matched those purchases with those that correlate to that of a pregnant women. It gets better, they don't even need you to buy 'What to Expect When You're Expecting' (book or DVD) – they can tell you are pregnant by correlations that we would not immediately associate with being pregnant – like buying scentless cosmetics. Tick off a couple of purchases in the right boxes, be from the right demographic and bam - they can even tell you when you are likely to be due…

And that is the other big theme in this book – bizarre correlations. Richard Feynman said somewhere that he is a Babylonian, rather than a Greek. By that he meant that the Greeks were always after 'proof', whereas the Babylonians were much more interested in mathematics that worked, whether they could prove why it worked or not. Proof is another word for 'cause' and this book is, in some ways, an extended mediation on that greatest of all First Year undergraduate lines 'correlation does not equal causation' – except this book shows that often understanding that two things are correlated, without having a clue as to why they might be, is actually enough. It doesn't equal causation, but it does mean a certain percentage of men on their way home are likely to buy a six pack of beer and nappies so the supermarket isn't going to have these items beside each other in the store or on special at the same time. You are being manipulated, but in ways you can never guess.

Now, there are incredible dangers to all of this big data stuff. Not the least that we now live in what is a total surveillance society. I read recently that a woman's Fitbit data was used against her in a court case. She was suing her employer for additional compensation for a workplace accident she was involved in because she had been highly active before an injury – but the insurance company was able to show, from her Fitbit data, that she hadn't been nearly as active as she had said she had been. The new Apple Watch will constantly measure and record your heart beat. The iPhone already records everything you do, the number of steps you take, where you are at any given time. It is actually also easy to also link that to who is near you at that time.

The next time your insurance premium goes up is it because they have guessed you fit within a certain demographic, or is it because they have access to more health information about you than your family doctor?

Eventually someone is going to be asked, “So, what were you doing at 12.37pm on 23 March 2016?” “And what about at 3.17pm on 9 June 2016?” And there will be an entire other list of dates and times. The punch-line being that you won’t have a clue what you were doing on those dates and times, but the person asking the questions will. They will know you were having coffee in various cafes about town and they will also know that Jack (aka The Butcher) Badass was in those cafes at those same times – how do you explain that, you international terrorist in waiting?

Sound far-fetched? Tell that to some of the poor bastards locked up and tortured in Guantanamo Bay for over a decade.

As is repeatedly said in this book, the Stasi were bad, but they didn’t have half the information we gather on people now as a matter of course.

I attended a lecture once on Bourdieu. Bourdieu was particularly interested in using data to understand how groups of people within society differentiated themselves from other groups of people. The person giving the lecture mentioned big data and how sociologists would give their eyeteeth to get their hands on some of it. Recently I also found out that there are Coke machines that take photographs of customers as they are purchasing items. The photograph can tell not only the time and date of purchase, but also the gender and race of the customer, not by someone in HQ having to look, but via an algorithm. Like I said, this data is a sociologist dream, but it is also mostly inaccessible – locked away as ‘commercial in confidence’. And in part, this is fair enough, there is an interesting discussion in this book on how data that might otherwise look like it has been de-identified can still be used to identify people. Again, this is based on unexpected correlations – you know, who’d have thought that there could be only one teacher in Milwaukie married to a Cuban-Chinese who likes Thai food and sings opera…

If there was one thing I didn’t like about this book it was its habit of making annoying puns, for example, “When the floor is datafied, there is no ceiling to its possible uses”. Hmm. And there are quite a few of these. I will spare you the others.

Years ago I read Predictably Irrational and thought after I’d finished, ‘God, advertisers must use these ideas all the time to get us to buy crap we don’t need’. This book shows how they actually do go about doing exactly that. I don’t for a minute believe there is any way we can go back to a small data world, and no technology is all good or all bad – but every technology changes the world, each has unintended consequences. We need to be aware of the fact, because some of the unintended consequences of big data are terrifying. We will be told there is nothing we can do about these consequences, that the benefits far outweigh the dangers – but the ironic consequence of this stuff is that it allows corporations to atomise us more than we have been already, it encourages us to think of ourselves as even more distinct and separate – as customers rather than citizens. There is a real price to be paid when we forsake our group rights – we will inevitably learn that price as time goes by and as big data becomes ever more invasive, ever more predictive and ever more manipulative.

Aaron Thibeault says

*A full executive summary of this book is available here: http://newbooksinbrief.com/2013/03/21...

The main argument: Statistical information, or data, has long been recognized to be a potentially rich and valuable source of knowledge. Until recently, however, our ability to render phenomena and events in a quantified format, store this information, and analyze it has been severely limited. With the rise of the digital age, though, these limitations are quickly being eroded. To begin with, digital devices that record our movements and communications, and digital sensors that record the behavior of inanimate objects and systems have become widespread and are proliferating wildly. What's more, the cost of storing this information on computer servers is getting cheaper and cheaper, thus allowing us to keep much more of it than ever before. Finally, increasingly sophisticated computer algorithms are allowing us to analyze this information more deeply than ever, and are revealing interesting (and often counter-intuitive) relationships that would never have been possible previously. The increasing datification of the world, and the insights that this is bringing us, may be thought of as one grand phenomenon, and it has a name: Big Data.

The insights that are emerging out of big data are spread out over many areas, and are already impacting several aspects of society. To begin with, big data is helping established businesses to run more efficiently and safely. For example, big data is being used to streamline assembly lines and also to catch quality control problems in the factory. But the benefits of big data go well beyond the factory. For example, the courier company UPS has used big data to help it map out more efficient trucking routes. The resulting improvements have allowed UPS to shave 30 million miles and 3 million gallons of fuel per year from their routes (loc. 1352). The more efficient trucking routes have also led to less traffic accidents. Meanwhile, car companies are beginning to use data from sensors in automobiles to understand which parts are causing problems, and also to understand where and why accidents are happening, so that they may be lessened.

In addition to helping already established businesses, big data is also allowing for new business opportunities that were never possible before. For example, the business prodigy Oren Etzioni used big data to set up a business called Farecast that predicts the cost of airfare tickets. When his business was bought by Microsoft for $110 million, Etzioni used big data again to set up a related business that predicts the cost of all manner of consumer goods. His very profitable business, Decide.com, saves consumers on average $100 per product (loc. 1867).

Outside of the business world, big data is also being used by governments to help reduce costs and make society safer. For example, in 2009 Google was able to apply big data to search terms to help identify how the H1N1 virus was spreading through communities in real time. This method of tracking disease pandemics holds great promise for allowing public health organizations to know when pandemics are beginning, and also to keep better track of how they are unfolding, in order that they may better contain them. In addition, big data is being used to help identify where potentially dangerous infrastructural problems are occurring, and also to identify trouble spots for fire hazards, in order that they may be addressed.

Big data also has significant potential uses in health care. Indeed, our increasing ability to monitor and record everything from our vital signs to the health of our systems to our individual genomes promises to inaugurate an age of personalized medicine that will allow doctors to more easily diagnose our ailments and tailor treatments to our individual bodies.

While big data may already be bringing us impressive benefits, Viktor Mayer-Schonberger and Kenneth Cukier argue that the bulk of the benefits are yet to come. Indeed, for the authors, businesses and

governments are only just now waking up to the incredible potential of Big Data. And as they direct more attention to recording and analyzing data streams, the potential uses of the information will only multiply.

On the negative side, big data also carries substantial potential dangers. Most notably, as more and more information about us is recorded, kept and used, our privacy is increasingly threatened. For the authors, a good deal of oversight will be needed in order to ensure that the potential abuses of big data are curbed.

The book is well written and represents a fine overview of the present and future of big data. Also, the authors do well to raise important big-picture issues related to the phenomena, though the potential impacts of big data (both positive and negative) are occasionally overblown. All in all the book is a good introduction to an important and interesting topic. A full executive summary of the book is available here: http://newbooksinbrief.com/2013/03/21... A podcast discussion of the book will be available soon.

Rob Kitchin says

In 2008 the term ‘big data’ was barely in use. Five years later and it has become latest ICT-related buzzword, used to refer to the recent surge in the generation of huge quantities of diverse and dynamic data produced by social media, transactions and interactions across the internet, sensor and camera networks, a myriad software-enabled devices, scientific equipment, etc. Mayer-Schonberger and Cukier’s book aims to provide an initial survey and analysis of the big data phenomena and what they call datafication; the process of transforming all things under the sun into data from which value can be extracted. They argue that a data revolution is underway, with the nature of data production and analysis undergoing a paradigm shift in three ways. First, the volume of data being produced is being radically transformed, with a move away from sampling to try and capture entire populations (n=all). Second, by being exhaustive in scope, it is possible to embrace the messiness of data rather than seeking exactitude (as required, along with randomness, in sampling); as they put it “more trumps better”. Third, that the types of questions asked changes from why (causation) to what (correlation): “We don’t always need to know the cause of a phenomena; rather, we can let the data speak for itself.” In other words, the traditional deductive, hypothesis-led mode of analysis is replaced by an inductive approach wherein analytics examine the data for all meaningful patterns, rather than testing for particular relationships. This third shift, they argue, also means that there is no longer the need for domain-specific expertise. As such, the era of big data is producing massive, exhaustive, messy datasets that can be mined for insightful information that can be used to identify relationships within the data that can be capitalised upon, such as using the vast quantities of data produced by a supermarket chain about consumers and their transactions to identify patterns of purchases which can then be used to tailor marketing strategies and increase turn-over.

Mayer-Schonberger and Cukier are right that there is a data revolution underway and they provide an initial overview of the big data phenomena. However, their analysis is weak in a number of respects. First, it ignores completely emerging debates about the kind of empiricism and data dredging they describe, which are deeply problematic in all kinds of ways, and the data-driven science being advocated by scientists. No scientist or analyst worth their salt believes that data simply speak for themselves free of theory. Second, the account is quite sketchy as to how analysts can make sense of big data and the new analytical techniques that are being developed. There is a science to big data in terms of devising the algorithms employed within machine learning and other big data analytics, yet the reader gets no real insight into how these work. Third, the book avoids tackling the deep and difficult epistemological questions that arise when a paradigm shift occurs. The book is clearly targeted at a non-academic audience, but nevertheless a grounded discussion of the philosophy of data and science in the era of big data is merited when such grandiose claims are being

made. Fourth, they rightly acknowledge that big data raises all kinds of ethical issues, but their analysis lacks depth and critique and pushes a business-friendly, market-orientated line about self-regulation without adequately setting out the pros and cons of such a strategy. Finally, the text suffers from being overly repetitious and the referencing is dreadful: it would be possible to condense the entire book into just a couple of chapters and not lose any of the argument, and whilst there are notes in back of book there are no numbers in the text to link to them. Overall then, whilst the book provides an initial text about big data and does include some interesting and useful nuggets, the analysis in general is narrow and weak, and it seems more about championing an emerging ICT market than providing a thorough, critical overview of the nature of big data and its implications and consequences.

Lars says

Having been on a big data/statistics binge recently, I can't help but be struck by the similarities in approach and execution between the major titles on the subject. Crack open any of these books and the authors are sure to regale you with the torrid tales of Billy Beane and his baseball Sabermetricians, Target's premature targeting of expectant mothers, and lest we forget - fawning references to the zany whizzes over at Google.

Still, the subject is so intensely fascinating that it doesn't matter much. My fascination, in no small part, comes from my belief that the big data approaches made possible by machine learning and distributed processing are a key component of where we are headed as a society. Today, algorithms can be used to translate languages (albeit it imperfect) and predict everything from crime hotspots to default rates, using data that seems almost unrelated to the matter at hand. The ingenuity and mathematical sophistication of the people who create these algorithms is awe-inspiring.

In covering them, "Big Data" does worse than most. It reads like a checklist of plane fodder tropes. Neologisms ("Algorithmists") abound, as do catchphrases ("N = All", "Big Data mindset"). The book lacks the practitioner's viewpoint of The Signal and the Noise: Why So Many Predictions Fail - But Some Don't. Unlike Automate This: How Algorithms Came to Rule Our World, it gently sidesteps the close but oh so uncomfortable issue of technological unemployment. Even more egregiously, it does this in favor of pages upon pages arguing against the distant straw man of Minority Report-style crime prediction. Throughout the book, one gets the sense that the authors and/or the editor had a vague desire of creating a manifesto for the new age of Big Data, but not the balls to go all out in the endeavour. Mashed together with the fluffy and excited tech journalism that takes up the majority of the book, the end result feels intellectually lazy and padded to the extreme.

I would not recommend this book.

Troy Blackford says

This is a fascinating exploration of the topic of what happens when the sample size of a given field of study reaches the point where n=all. That is to say, what happens when we have data for all of a given variable, rather than just taking a sampling and extrapolating out findings. The authors compare the difference to what happened to visual representation when you could make a number of photographs in one second - it wasn't just a quality-based leap like that of the jump from realistic artwork to photography, but a leap into another realm - the ability to capture and display motion pictures.

The same type of leap is happening with data - the preponderance of information is allowing our understanding if that data to become something else, something more than the sum of its parts. This book explores everything from Google tracking flu outbreaks via search terms to New York City officials determining which buildings have been illegally modified into being major fire risks - all using data collected for other means. Along the way you will be introduced to pioneers in the use of data, going all the way back to a 19th century naval desk clerk who used forgotten weather records to deduce the best trade routes, to modern experts who did everything from make the best possible spell checker algorithms out of peoples' mistakes to blowing the lid off the best machine translators of the day using piles and piles of scanned books.

This was an awesome read and I recommend it to anyone interested in the topic of data and how it is used to enhance and define our modern lives.

Mal Warwick says

Shocked by the NSA revelations? You don't know the whole story

While Edward Snowden bounces from one temporary refuge to another in search of safe harbor from the long arms of the U.S. government, the American public is starting to wake up to the reality of Big Data. The National Security Agency, long one of the pioneers in this burgeoning but little-appreciated field, has been teaching us -- or, rather, Snowden, The Guardian, and the Washington Post have been teaching us -- about the power that resides in gargantuan masses of data. Now here come Viktor Mayer-Schoeneberger and Kenneth Cukier with a new book that goes far beyond the headlines about espionage and invasion of privacy to give us an eminently readable, well-organized overview of Big Data's origins, its characteristics, and its potential for both good and evil.

When we think of Big Data, we, or at least most of us, think of computers. However, the authors persuade us that the fundamentals of Big Data were laid down more than a century before the invention of the microprocessor. They point to a legendary American seaman named Matthew Maury. In the middle of the 19th Century, after 16 years of effort, Maury published a book based on 1.2 billion data points gleaned from old ships' logs stored by the Navy that dramatically reduced the distances (and, hence, the time elapsed) in ocean voyages by both military and commercial ships. Maury used facts derived from decades of mariners' observations to dispel the myths, legends, superstitions, and rumors that had long caused ocean-voyaging ships to pursue roundabout courses. Not so incidentally, Maury's work also facilitated the laying of the first transatlantic telegraph cable.

If not the first, this was certainly an early application of Big Data, which the authors describe as follows: "big data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value, in ways that change markets, organizations, the relationship between citizens and governments, and more." For example, if Maury had had available only a fraction of the old ships' logs he found in the naval archives, his task would have been impractical, since each individual log doubtless included small errors (and an occasional big one). Only by amassing a huge store of data did those errors cancel out one another.

Now, in the Digital Age, the volumes of data that can be harnessed are, at times, literally astronomical. "Google processes more than 24 petabytes of data per day, a volume that is thousands of times the quantity of all printed material in the U.S. Library of Congress." AT&T transfers about 30 petabytes of data through

its networks each day. Twenty-four or 30 of something doesn't sound like much, unless you understand that a megabyte is a million bytes, a gigabyte is a billion bytes, a terabyte is 1,000 times the size of a gigabyte, and a petabyte is 1,000 times the size of a terabyte. That's 1,000,000,000,000,000 bytes. That's a lot of data! But even that's only a tiny slice of all the data now stored in the world, "estimated to be around 1,200 exabytes." And an exabyte (I'm sure you're dying to know) is the equivalent of 1,000 petabytes. So, 1,200 petabytes could also be stated as 1.2 zettabytes, with a zettabyte equal to 1,000 petabytes, and I'll bet that not one person in a million has ever heard of a zettabyte before. Had you?

All of which should make clear that when we talk about Big Data today, we're talking about really, really big numbers -- so big, in fact, that almost no matter how messy or inaccurate the data might be, it's usually possible to draw useful, on-target insights from analyzing it. That's what's different about Big Data -- and that's why the phenomenon is bound to change the way we think about the world.

We live in a society obsessed with causality. We often care more about why something happened than about what it was that happened. And in a world where Big Data looms larger and larger all the time, we'll have to get used to not knowing -- or even caring much -- why things happen.

"At its core," write Mayer-Schoeneberger and Cukier, "big data is about predictions. Though it is described as part of the branch of computer science called artificial intelligence, and more specifically, an area called machine learning, this characterization is misleading. Big data is not about trying to 'teach' a computer to 'think' like humans. Instead, it's about applying math to huge quantities of data in order to infer probabilities: the likelihood that an email message is spam; that the typed letters 'teh' are supposed to be 'the'; that the trajectory and velocity of a person jaywalking mean he'll make it across the street in time [so that] the self-driving car need only slow slightly."

The authors refer to data as "the oil of the information economy," predicting that, as it flows into all the nooks and crannies of our society, it will bring about "three major shifts of mindset that are interlinked and hence reinforce one another." First among these is our ever-growing ability to analyze inconceivably large amounts of data and not have to settle for sampling. Second, we'll come to accept the inevitable messiness in huge stores of data and learn not to insist on precision in reporting. Third, and last, we'll get used to accepting correlations rather than causality. "The ideal of identifying causal mechanisms is a self-congratulatory illusion; big data overturns this," the authors assert.

If you want to understand this increasingly important aspect of contemporary life, I suggest you read Big Data.

Viktor Mayer-Schoeneberger and Kenneth Cukier come to the task of writing this book with unbeatable credentials. Mayer-Schoeneberger is Professor of Internet Governance at Oxford University, and Kenneth Cukier is Data Editor at The Economist.
