BIG DATA

BIG DATA

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

VIKTOR MAYER-SCHÖNBERGER
AND KENNETH CUKIER

JOHN MURRAY

First published in Great Britain in 2013 by John Murray (Publishers)
An Hachette UK Company

1

A CIP catalogue record for this title is available from the British Library

Hardback ISBN 978-1-84854-790-2
Trade Paperback ISBN 978-1-84854-791-9
Ebook ISBN 978-1-84854-793-3

Printed and bound by Clays Ltd, St Ives plc

John Murray policy is to use papers that are natural, renewable and recyclable products and made from wood grown in sustainable forests. The logging and manufacturing processes are expected to conform to the environmental regulations of the country of origin.

John Murray (Publishers)
338 Euston Road
London NW1 3BH

www.johnmurray.co.uk

To B and V

V.M.S.

To my parents

K.N.C.

CONTENTS

BIG DATA

1

NOW

IN 2009 A NEW FLU virus was discovered. Combining elements of the viruses that cause bird flu and swine flu, this new strain, dubbed H1N1, spread quickly. Within weeks, public health agencies around the world feared a terrible pandemic was under way. Some commentators warned of an outbreak on the scale of the 1918 Spanish flu that had infected half a billion people and killed tens of millions. Worse, no vaccine against the new virus was readily available. The only hope public health authorities had was to slow its spread. But to do that, they needed to know where it already was.

In the United States, the Centers for Disease Control and Prevention (CDC) requested that doctors inform them of new flu cases. Yet the picture of the pandemic that emerged was always a week or two out of date. People might feel sick for days but wait before consulting a doctor. Relaying the information back to the central organizations took time, and the CDC only tabulated the numbers once a week. With a rapidly spreading disease, a two-week lag is an eternity. This delay completely blinded public health agencies at the most crucial moments.

As it happened, a few weeks before the H1N1 virus made headlines, engineers at the Internet giant Google published a remarkable paper in the scientific journal *Nature*. It created a splash among health officials and computer scientists but was otherwise overlooked. The authors explained how Google could "predict" the spread of the winter flu in the United States, not just nationally, but down to specific regions and even states. The company could achieve this by looking at what people were searching for on the Internet. Since Google receives more than three billion search queries every day and saves them all, it had plenty of data to work with.

Google took the 50 million most common search terms that Americans type and compared the list with CDC data on the spread of seasonal flu between 2003 and 2008. The idea was to identify areas infected by the flu virus by what people searched for on the Internet. Others had tried to do this with Internet search terms, but no one else had as much data, processing power, and statistical know-how as Google.

While the Googlers guessed that the searches might be aimed at getting flu information — typing phrases like "medicine for cough and fever" — that wasn't the point: they didn't know, and they designed a system that didn't care. All their system did was look for correlations between the frequency of certain search queries and the spread of the flu over time and space. In total, they processed a staggering 450 million different mathematical models in order to test the search terms, comparing their predictions against actual flu cases from the CDC in 2007 and 2008. And they struck gold: their software found a combination of 45 search terms that, when used together in a mathematical model, had a strong correlation between their prediction and the official figures nationwide. Like the CDC, they could tell where the flu had spread, but unlike the CDC they could tell it in near real time, not a week or two after the fact.

Thus when the H1N1 crisis struck in 2009, Google's system proved to be a more useful and timely indicator than government statistics with their natural reporting lags. Public health officials were armed with valuable information.

Strikingly, Google's method does not involve distributing mouth swabs or contacting physicians' offices. Instead, it is built on "big data" — the ability of society to harness information in novel ways to produce useful insights or goods and services of significant value.

With it, by the time the next pandemic comes around, the world will have a better tool at its disposal to predict and thus prevent its spread.

Public health is only one area where big data is making a big difference. Entire business sectors are being reshaped by big data as well. Buying airplane tickets is a good example.

In 2003 Oren Etzioni needed to fly from Seattle to Los Angeles for his younger brother's wedding. Months before the big day, he went online and bought a plane ticket, believing that the earlier you book, the less you pay. On the flight, curiosity got the better of him and he asked the fellow in the next seat how much his ticket had cost and when he had bought it. The man turned out to have paid considerably less than Etzioni, even though he had purchased the ticket much more recently. Infuriated, Etzioni asked another passenger and then another. Most had paid less.

For most of us, the sense of economic betrayal would have dissipated by the time we closed our tray tables and put our seats in the full, upright, and locked position. But Etzioni is one of America's foremost computer scientists. He sees the world as a series of big-data problems — ones that he can solve. And he has been mastering them since he graduated from Harvard in 1986 as its first undergrad to major in computer science.

From his perch at the University of Washington, he started a slew of big-data companies before the term "big data" became known. He helped build one of the Web's first search engines, MetaCrawler, which was launched in 1994 and snapped up by InfoSpace, then a major online property. He co-founded Netbot, the first major comparison-shopping website, which he sold to Excite. His startup for extracting meaning from text documents, called ClearForest, was later acquired by Reuters.

Back on terra firma, Etzioni was determined to figure out a way for people to know if a ticket price they see online is a good deal or not. An airplane seat is a commodity: each one is basically indistinguishable from others on the same flight. Yet the prices vary wildly, based on a myriad of factors that are mostly known only by the airlines themselves.

Etzioni concluded that he didn't need to decrypt the rhyme or reason for the price differences. Instead, he simply had to predict whether the price being shown was likely to increase or decrease in the future. That is possible, if not easy, to do. All it requires is analyzing all the ticket sales for a given route and examining the prices paid relative to the number of days before the departure.

If the average price of a ticket tended to decrease, it would make sense to wait and buy the ticket later. If the average price usually increased, the system would recommend buying the ticket right away at the price shown. In other words, what was needed was a souped-up version of the informal survey Etzioni conducted at 30,000 feet. To be sure, it was yet another massive computer science problem. But again, it was one he could solve. So he set to work.

Using a sample of 12,000 price observations that was obtained by "scraping" information from a travel website over a 41-day period, Etzioni created a predictive model that handed its simulated passengers a tidy savings. The model had no understanding of *why,* only *what*. That is, it didn't know any of the variables that go into airline pricing decisions, such as number of seats that remained unsold, seasonality, or whether some sort of magical Saturday-night-stay might reduce the fare. It based its prediction on what it did know: probabilities gleaned from the data about other flights. "To buy or not to buy, that is the question," Etzioni mused. Fittingly, he named the research project Hamlet.

The little project evolved into a venture capital–backed startup called Farecast. By predicting whether the price of an airline ticket was likely to go up or down, and by how much, Farecast empowered consumers to choose when to click the "buy" button. It armed them with information to which they had never had access before. Upholding the virtue of transparency against itself, Farecast even scored the degree of confidence it had in its own predictions and presented that information to users too.

To work, the system needed lots of data. To improve its performance, Etzioni got his hands on one of the industry's flight reservation databases. With that information, the system could make predictions based on every seat on every flight for most routes in American commercial aviation over the course of a year. Farecast was now crunching nearly 200 billion flight-price records to make its predictions. In so doing, it was saving consumers a bundle.

With his sandy brown hair, toothy grin, and cherubic good looks, Etzioni hardly seemed like the sort of person who would deny the airline industry millions of dollars of potential revenue. In fact, he set his sights on doing even more than that. By 2008 he was planning to apply the method to other goods like hotel rooms, concert tickets, and used cars: anything with little product differentiation, a high degree of price variation, and tons of data. But before he could hatch his plans, Microsoft came knocking on his door, snapped up Farecast for around $110 million, and integrated it into the Bing search engine. By 2012 the system was making the correct call 75 percent of the time and saving travelers, on average, $50 per ticket.

Farecast is the epitome of a big-data company and an example of where the world is headed. Etzioni couldn't have built the company five or ten years earlier. "It would have been impossible," he says. The amount of computing power and storage he needed was too expensive. But although changes in technology have been a critical factor making it possible, something more important changed too, something subtle. There was a shift in mindset about how data could be used.

Data was no longer regarded as static or stale, whose usefulness was finished once the purpose for which it was collected was achieved, such as after the plane landed (or in Google's case, once a search query had been processed). Rather, data became a raw material of business, a vital economic input, used to create a new form of economic value. In fact, with the right mindset, data can be cleverly reused to become a fountain of innovation and new services. The data can reveal secrets to those with the humility, the willingness, and the tools to listen.

Letting the data speak

The fruits of the information society are easy to see, with a cellphone in every pocket, a computer in every backpack, and big information technology systems in back offices everywhere. But less noticeable is the information itself. Half a century after computers entered mainstream society, the data has begun to accumulate to the point where something new and special is taking place. Not only is the world awash with more information than ever before, but that information is growing faster. The change of scale has led to a change of state. The quantitative change has led to a qualitative one. The sciences like astronomy and genomics, which first experienced the explosion in the 2000s, coined the term "big data." The concept is now migrating to all areas of human endeavor.

There is no rigorous definition of big data. Initially the idea was that the volume of information had grown so large that the quantity being examined no longer fit into the memory that computers use for processing, so engineers needed to revamp the tools they used for analyzing it all. That is the origin of new processing technologies like Google's MapReduce and its open-source equivalent, Hadoop, which came out of Yahoo. These let one manage far larger quantities of data than before, and the data — importantly — need not be placed in tidy rows or classic database tables. Other data-crunching technologies that dispense with the rigid hierarchies and homogeneity of yore are also on the horizon. At the same time, because Internet companies could collect vast troves of data and had a burning financial incentive to make sense of them, they became the leading users of the latest processing technologies, superseding offline companies that had, in some cases, decades more experience.

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, organizations, the relationship between citizens and governments, and more.

But this is just the start. The era of big data challenges the way we live and interact with the world. Most strikingly, society will need to shed some of its obsession for causality in exchange for simple correlations: not knowing *why* but only *what*. This overturns centuries of established practices and challenges our most basic understanding of how to make decisions and comprehend reality.

Big data marks the beginning of a major transformation. Like so many new technologies, big data will surely become a victim of Silicon Valley's notorious hype cycle: after being feted on the cover of magazines and at industry conferences, the trend will be dismissed and many of the data-smitten startups will flounder. But both the infatuation and the damnation profoundly misunderstand the importance of what is taking place. Just as the telescope enabled us to comprehend the universe and the microscope allowed us to understand germs, the new techniques for collecting and analyzing huge bodies of data will help us make sense of our world in ways we are just starting to appreciate. In this book we are not so much big data's evangelists, but merely its messengers. And, again, the real revolution is not in the machines that calculate data but in data itself and how we use it.

To appreciate the degree to which an information revolution is already under way, consider trends from across the spectrum of society. Our digital universe is constantly expanding. Take astronomy. When the Sloan Digital Sky Survey began in 2000, its telescope in New Mexico collected more data in its first few weeks than had been amassed in the entire history of astronomy. By 2010 the survey's archive teemed with a whopping 140 terabytes of information. But a successor, the Large Synoptic Survey Telescope in Chile, due to come on stream in 2016, will acquire that quantity of data every five days.

Such astronomical quantities are found closer to home as well. When scientists first decoded the human genome in 2003, it took them a decade of intensive work to sequence the three billion base pairs. Now, a decade later, a single facility can sequence that much DNA in a day. In finance, about seven billion shares change hands every day on U.S. equity markets, of which around two-thirds is traded by computer algorithms based on mathematical models that crunch mountains of data to predict gains while trying to reduce risk.

Internet companies have been particularly swamped. 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. Facebook, a company that didn't exist a decade ago, gets more than 10 million new photos uploaded every hour. Facebook members click a "like" button or leave a comment nearly three billion times per day, creating a digital trail that the company can mine to learn about users' preferences. Meanwhile, the 800 million monthly users of Google's YouTube service upload over an hour of video every second. The number of messages on Twitter grows at around 200 percent a year and by 2012 had exceeded 400 million tweets a day.

From the sciences to healthcare, from banking to the Internet, the sectors may be diverse yet together they tell a similar story: the amount of data in the world is growing fast, outstripping not just our machines but our imaginations.

Many people have tried to put an actual figure on the quantity of information that surrounds us and to calculate how fast it grows. They've had varying degrees of success because they've measured different things. One of the more comprehensive studies was done by Martin Hilbert of the University of Southern California's Annenberg School for Communication and Journalism. He has striven to put a figure on everything that has been produced, stored, and communicated. That would include not only books, paintings, emails, photographs, music, and video (analog and digital), but video games, phone calls, even car navigation systems and letters sent through the mail. He also included broadcast media like television and radio, based on audience reach.

By Hilbert's reckoning, more than 300 exabytes of stored data existed in 2007. To understand what this means in slightly more human terms, think of it like this. A full-length feature film in digital form can be compressed into a one gigabyte file. An exabyte is one billion gigabytes. In short, it's a lot. Interestingly, in 2007 only about 7 percent of the data was analog (paper, books, photographic prints, and so on). The rest was digital. But not long ago the picture looked very different. Though the ideas of the "information revolution" and "digital age" have been around since the 1960s, they have only just become a reality by some measures. As recently as the year 2000, only a quarter of the stored information in the world was digital. The other three-quarters were on paper, film, vinyl LP records, magnetic cassette tapes, and the like.

The mass of digital information then was not much — a humbling thought for those who have been surfing the Web and buying books online for a long time. (In fact, in 1986 around 40 percent of the world's general-purpose computing power took the form of pocket calculators, which represented more processing power than all personal computers at the time.) But because digital data expands so quickly — doubling a little more than every three years, according to Hilbert — the situation quickly inverted itself. Analog information, in contrast, hardly grows at all. So in 2013 the amount of stored information in the world is estimated to be around 1,200 exabytes, of which less than 2 percent is non-digital.

There is no good way to think about what this size of data means. If it were all printed in books, they would cover the entire surface of the United States some 52 layers thick. If it were placed on CD-ROMs and stacked up, they would stretch to the moon in five separate piles. In the third century B.C., as Ptolemy II of Egypt strove to store a copy of every written work, the great Library of Alexandria represented the sum of all knowledge in the world. The digital deluge now sweeping the globe is the equivalent of giving every person living on Earth today 320 times as much information as is estimated to have been stored in the Library of Alexandria.

Things really are speeding up. The amount of stored information grows four times faster than the world economy, while the processing power of computers grows nine times faster. Little wonder that people complain of information overload. Everyone is whiplashed by the changes.

Take the long view, by comparing the current data deluge with an earlier information revolution, that of the Gutenberg printing press, which was invented around 1439. In the fifty years from 1453 to 1503 about eight million books were printed, according to the historian Elizabeth Eisenstein. This is considered to be more than all the scribes of Europe had produced since the founding of Constantinople some 1,200 years earlier. In other words, it took 50 years for the stock of information to roughly double in Europe, compared with around every three years today.

What does this increase mean? Peter Norvig, an artificial intelligence expert at Google, likes to think about it with an analogy to images. First, he asks us to consider the iconic horse from the cave paintings in Lascaux, France, which date to the Paleolithic Era some 17,000 years ago. Then think of a photograph of a horse — or better, the dabs of Pablo Picasso, which do not look much dissimilar to the cave paintings. In fact, when Picasso was shown the Lascaux images he quipped that, since then, "We have invented nothing."

Picasso's words were true on one level but not on another. Recall that photograph of the horse. Where it took a long time to draw a picture of a horse, now a representation of one could be made much faster with photography. That is a change, but it may not be the most essential, since it is still fundamentally the same: an image of a horse. Yet now, Norvig implores, consider capturing the image of a horse and speeding it up to 24 frames per second. Now, the quantitative change has produced a qualitative change. A movie is fundamentally different from a frozen photograph. It's the same with big data: by changing the amount, we change the essence.

Consider an analogy from nanotechnology — where things get smaller, not bigger. The principle behind nanotechnology is that when you get to the molecular level, the physical properties can change. Knowing those new characteristics means you can devise materials to do things that could not be done before. At the nanoscale, for example, more flexible metals and stretchable ceramics are possible. Conversely, when we increase the scale of the data that we work with, we can do new things that weren't possible when we just worked with smaller amounts.

Sometimes the constraints that we live with, and presume are the same for everything, are really only functions of the scale in which we operate. Take a third analogy, again from the sciences. For humans, the single most important physical law is gravity: it reigns over all that we do. But for tiny insects, gravity is mostly immaterial. For some, like water striders, the operative law of the physical universe is surface tension, which allows them to walk across a pond without falling in.

With information, as with physics, size matters. Hence, Google is able to identify the prevalence of the flu just about as well as official data based on actual patient visits to the doctor. It can do this by combing through hundreds of billions of search terms—and it can produce an answer in near real time, far faster than official sources. Likewise, Etzioni's Farecast can predict the price volatility of an airplane ticket and thus shift substantial economic power into the hands of consumers. But both can do so well only by analyzing hundreds of billions of data points.

These two examples show the scientific and societal importance of big data as well as the degree to which big data can become a source of economic value. They mark two ways in which the world of big data is poised to shake up everything from businesses and the sciences to healthcare, government, education, economics, the humanities, and every other aspect of society.

Although we are only at the dawn of big data, we rely on it daily. Spam filters are designed to automatically adapt as the types of junk email change: the software couldn't be programmed to know to block "via6ra" or its infinity of variants. Dating sites pair up couples on the basis of how their numerous attributes correlate with those of successful previous matches. The "autocorrect" feature in smartphones tracks our actions and adds new words to its spelling dictionary based on what we type. Yet these uses are just the start. From cars that can detect when to swerve or brake to IBM's Watson computer beating humans on the game show *Jeopardy!,* the approach will revamp many aspects of the world in which we live.

At its core, 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—the self-driving car need only slow slightly. The key is that these systems perform well because they are fed with lots of data on which to base their predictions. Moreover, the systems are built to improve themselves over time, by keeping a tab on what are the best signals and patterns to look for as more data is fed in.

In the future—and sooner than we may think—many aspects of our world will be augmented or replaced by computer systems that today are the sole purview of human judgment. Not just driving or matchmaking, but even more complex tasks. After all, Amazon can recommend the ideal book, Google can rank the most relevant website, Facebook knows our likes, and LinkedIn divines whom we know. The same technologies will be applied to diagnosing illnesses, recommending treatments, perhaps even identifying "criminals" before one actually commits a crime. Just as the Internet radically changed the world by adding communications to computers, so too will big data change fundamental aspects of life by giving it a quantitative dimension it never had before.

More, messy, good enough

Big data will be a source of new economic value and innovation. But even more is at stake. Big data's ascendancy represents three shifts in the way we analyze information that transform how we understand and organize society.

The first shift is described in Chapter Two. In this new world we can analyze far more data. In some cases we can even process *all* of it relating to a particular phenomenon. Since the nineteenth century, society has depended on using samples when faced with large numbers. Yet the need for sampling is an artifact of a period of information scarcity, a product of the natural constraints on interacting with information in an analog era. Before the prevalence of high-performance digital technologies, we didn't recognize sampling as artificial fetters — we usually just took it for granted. Using all the data lets us see details we never could when we were limited to smaller quantities. Big data gives us an especially clear view of the granular: subcategories and submarkets that samples can't assess.

Looking at vastly more data also permits us to loosen up our desire for exactitude, the second shift, which we identify in Chapter Three. It's a tradeoff: with less error from sampling we can accept more measurement error. When our ability to measure is limited, we count only the most important things. Striving to get the exact number is appropriate. It is no use selling cattle if the buyer isn't sure whether there are 100 or only 80 in the herd. Until recently, all our digital tools were premised on exactitude: we assumed that database engines would retrieve the records that perfectly matched our query, much as spreadsheets tabulate the numbers in a column.

This type of thinking was a function of a "small data" environment: with so few things to measure, we had to treat what we did bother to quantify as precisely as possible. In some ways this is obvious: a small store may count the money in the cash register at the end of the night down to the penny, but we wouldn't — indeed couldn't — do the same for a country's gross domestic product. As scale increases, the number of inaccuracies increases as well.

Exactness requires carefully curated data. It may work for small quantities, and of course certain situations still require it: one either does or does not have enough money in the bank to write a check. But in return for using much more comprehensive datasets we can shed some of the rigid exactitude in a big-data world.

Often, big data is messy, varies in quality, and is distributed among countless servers around the world. With big data, we'll often be satisfied with a sense of general direction rather than knowing a phenomenon down to the inch, the penny, the atom. We don't give up on exactitude entirely; we only give up our devotion to it. What we lose in accuracy at the micro level we gain in insight at the macro level.

These two shifts lead to a third change, which we explain in Chapter Four: a move away from the age-old search for causality. As humans we have been conditioned to look for causes, even though searching for causality is often difficult and may lead us down the wrong paths. In a big-data world, by contrast, we won't have to be fixated on causality; instead we can discover patterns and correlations in the data that offer us novel and invaluable insights. The correlations may not tell us precisely *why* something is happening, but they alert us *that* it is happening.

And in many situations this is good enough. If millions of electronic medical records reveal that cancer sufferers who take a certain combination of aspirin and orange juice see their disease go into remission, then the exact cause for the improvement in health may be less important than the fact that they lived. Likewise, if we can save money by knowing the best time to buy a plane ticket without understanding the method behind airfare madness, that's good enough. Big data is about *what,* not *why.* We don't always need to know the cause of a phenomenon; rather, we can let data speak for itself.

Before big data, our analysis was usually limited to testing a small number of hypotheses that we defined well before we even collected the data. When we let the data speak, we can make connections that we had never thought existed. Hence, some hedge funds parse Twitter to predict the performance of the stock market. Amazon and Netflix base their product recommendations on a myriad of user interactions on their sites. Twitter, LinkedIn, and Facebook all map users' "social graph" of relationships to learn their preferences.

Of course, humans have been analyzing data for millennia. Writing was developed in ancient Mesopotamia because bureaucrats wanted an efficient tool to record and keep track of information. Since biblical times governments have held censuses to gather huge datasets on their citizenry, and for two hundred years actuaries have similarly collected large troves of data concerning the risks they hope to understand—or at least avoid.

Yet in the analog age collecting and analyzing such data was enormously costly and time-consuming. New questions often meant that the data had to be collected again and the analysis started afresh.

The big step toward managing data more efficiently came with the advent of digitization: making analog information readable by computers, which also makes it easier and cheaper to store and process. This advance improved efficiency dramatically. Information collection and analysis that once took years could now be done in days or even less. But little else changed. The people who analyzed the data were too often steeped in the analog paradigm of assuming that datasets had singular purposes to which their value was tied. Our very processes perpetuated this prejudice. As important as digitization was for enabling the shift to big data, the mere existence of computers did not make big data happen.

There's no good term to describe what's taking place now, but one that helps frame the changes is *datafication,* a concept that we introduce in Chapter Five. It refers to taking information about all things under the sun—including ones we never used to think of as information at all, such as a person's location, the vibrations of an engine, or the stress on a bridge—and transforming it into a data format to make it quantified. This allows us to use the information in new ways, such as in predictive analysis: detecting that an engine is prone to a breakdown based on the heat or vibrations that it produces. As a result, we can unlock the implicit, latent value of the information.

There is a treasure hunt under way, driven by the insights to be extracted from data and the dormant value that can be unleashed by a shift from causation to correlation. But it's not just one treasure. Every single dataset is likely to have some intrinsic, hidden, not yet unearthed value, and the race is on to discover and capture all of it.

Big data changes the nature of business, markets, and society, as we describe in Chapters Six and Seven. In the twentieth century, value shifted from physical infrastructure like land and factories to intangibles such as brands and intellectual property. That now is expanding to data, which is becoming a significant corporate asset, a vital economic input, and the foundation of new business models. It is the oil of the information economy. Though data is rarely recorded on corporate balance sheets, this is probably just a question of time.

Although some data-crunching techniques have been around for a while, in the past they were only available to spy agencies, research labs, and the world's biggest companies. After all, Walmart and Capital One pioneered the use of big data in retailing and banking and in so doing changed their industries. Now many of these tools have been democratized (although the data has not).

The effect on individuals may be the biggest shock of all. Specific area expertise matters less in a world where probability and correlation are paramount. In the movie *Moneyball,* baseball scouts were upstaged by statisticians when gut instinct gave way to sophisticated analytics. Similarly, subject-matter specialists will not go away, but they will have to contend with what the big-data analysis says. This will force an adjustment to traditional ideas of management, decision-making, human resources, and education.

Most of our institutions were established under the presumption that human decisions are based on information that is small, exact, and causal in nature. But the situation changes when the data is huge, can be processed quickly, and tolerates inexactitude. Moreover, because of the data's vast size, decisions may often be made not by humans but by machines. We consider the dark side of big data in Chapter Eight.

Society has millennia of experience in understanding and overseeing human behavior. But how do you regulate an algorithm? Early on in computing, policymakers recognized how the technology could be used to undermine privacy. Since then society has built up a body of rules to protect personal information. But in an age of big data, those laws constitute a largely useless Maginot Line. People willingly share information online—a central feature of the services, not a vulnerability to prevent.

Meanwhile the danger to us as individuals shifts from privacy to probability: algorithms will predict the likelihood that one will get a heart attack (and pay more for health insurance), default on a mortgage (and be denied a loan), or commit a crime (and perhaps get arrested in advance). It leads to an ethical consideration of the role of free will versus the dictatorship of data. Should individual volition trump big data, even if statistics argue otherwise? Just as the printing press prepared the ground for laws guaranteeing free speech — which didn't exist earlier because there was so little written expression to protect — the age of big data will require new rules to safeguard the sanctity of the individual.

In many ways, the way we control and handle data will have to change. We're entering a world of constant data-driven predictions where we may not be able to explain the reasons behind our decisions. What does it mean if a doctor cannot justify a medical intervention without asking the patient to defer to a black box, as the physician must do when relying on a big-data-driven diagnosis? Will the judicial system's standard of "probable cause" need to change to "probabilistic cause" — and if so, what are the implications of this for human freedom and dignity?

New principles are needed for the age of big data, which we lay out in Chapter Nine. Although they build upon the values that were developed and enshrined for the world of small data, it's not simply a matter of refreshing old rules for new circumstances, but recognizing the need for new principles altogether.

The benefits to society will be myriad, as big data becomes part of the solution to pressing global problems like addressing climate change, eradicating disease, and fostering good governance and economic development. But the big-data era also challenges us to become better prepared for the ways in which harnessing the technology will change our institutions and ourselves.

Big data marks an important step in humankind's quest to quantify and understand the world. A preponderance of things that could never be measured, stored, analyzed, and shared before is becoming datafied. Harnessing vast quantities of data rather than a small portion, and privileging more data of less exactitude, opens the door to new ways of understanding. It leads society to abandon its time-honored preference for causality, and in many instances tap the benefits of correlation.

The ideal of identifying causal mechanisms is a self-congratulatory illusion; big data overturns this. Yet again we are at a historical impasse where "god is dead." That is to say, the certainties that we believed in are once again changing. But this time they are being replaced, ironically, by better evidence. What role is left for intuition, faith, uncertainty, acting in contradiction of the evidence, and learning by experience? As the world shifts from causation to correlation, how can we pragmatically move forward without undermining the very foundations of society, humanity, and progress based on reason? This book intends to explain where we are, trace how we got here, and offer an urgently needed guide to the benefits and dangers that lie ahead.