Big Data: A Revolution That Will Transform How We Live, Work, and Think, Viktor Mayer-Schonberger, Kenner Cukier

Data has become a raw material of business, a vital economic input used to create a new form of economic value. 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. 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. Society will shed some of its obsession for causality in exchange for simple correlations: not know *why* but only *what*. In 2013 the amount of stored information in the world is estimated to be around 1200 exabytes, of which less than 2% is non-digital.

A movie is fundamentally different from a frozen photograph. It's the same with big data: by changing the amount, we change the essence. 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 is surface tension, which allows them to walk across a pond without falling in. **With information, as with physics, size matters.**

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. Big data's ascendency represents shifts in the way we use information that transforms how we understand and organize society. We can analyze far more data and it gives us a clear view of the granular: subcategories and submarkets that samples can't assess. What we lose in accuracy at the micro level we gain in insight at the macro level. We also move away from the age-old search for causality toward correlation. Correlations may not tell us precisely why something is happening, but they alert us that it is happening. We can make connections we had never thought existed.

Datafication 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. This allows us to use the information in new ways, such as predictive analysis: detecting that an engine is prone to a break-down based on the heat or vibrations that it produces. As a result, we can unlock the implicit, latent value of the information.

In the 20th century, value shifted from physical infrastructure like land and factories to intangibles such as brands and intellectual property. That is now expanding to **data,** which **is becoming a significant corporate asset.** Wal-Mart 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 digitized.

Specific area expertise matters less in a world where probability and correlation are paramount. I**t changed baseball.** In the movie *Moneyball,* baseball scouts were upstaged by statisticians when gut instinct gave way to sophisticated analytics. And when the data showed that stealing bases was inefficient, out went one of the most exciting, but least "productive" elements of the game. Similarly, subject-matter specialists will not go away, but they will have to contend with what big-data analysis says.

As social media is driven by people who willingly share information online, 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 society to abandon its time-honored preference for causality,** and in many instances tap the benefits of correlation.

Big data is about major shifts of mindset that are interlinked and reinforce one another. First is the ability to analyze vast amounts of data about a topic rather than be forced to settle for smaller sets. Second is a willingness to embrace data's real-world messiness more than exactitude. Third is a growing respect for correlations rather than a continuing quest for causality.

One aim of statistics is to confirm the richest findings using the smallest amount of data. Statisticians have shown that sampling precision improves most dramatically with randomness, not with increased sample size. This paved the way for a new approach to gathering information. Data using random samples could be collected at low cost and yet extrapolated with high accuracy to the whole. Random sampling has been a huge success and is the backbone of modern measurement, but achieving randomness is tricky. Systematic biases in the way the data is collected can lead to the extrapolated results being very wrong. And sampling quickly stops being useful when you want to drill deeper, to take a close look at some intriguing subcategory in the data. What works at the macro level falls apart in the micro. The **concept of sampling no longer makes as much sense** when we can harness large amounts of data.

Temporal data is gathered over time, and as it accumulates you get more and more insight into patterns. So, we can toss aside the shortcut of random sampling and aim for more comprehensive data instead. Doing so requires

affordable collecting along with ample processing and storage power. As the cost and complexity of all these pieces of the puzzle have declined dramatically, what was previously the purview of just the biggest companies is now possible for most.

One of the areas being most dramatically shaken up by N=all is the **social sciences. They have lost their monopoly on making sense of empirical social data,** as big-data analysis replaces the highly skilled survey specialists of the past. The social science disciplines largely relied on sampling studies and questionnaires. But when the data is collected passively while people do what they normally do anyway, the old biases associated with sampling and questionnaires disappear as does the need to sample. Reaching for a random sample in the age of big data is like clutching at a horse whip in the era of the motor car.

One discovery of big data is that the quality of a social network degrades quickly as people with links outside their immediate community are taken off the network. Its structure quickly buckles. People with lots of close friends are far less important to the stability of the network structure than the one with ties to more distant people. It suggests that **there is a premium on diversity within a group and in society at large.**

By the 19th century, France—then the world's leading scientific nation—had developed a system of precisely defined units of measurement to capture space and time and had begun to get other nations to adopt the same standards—the metric system. 5% of all digital data is "structured"—that is, in a form that fits neatly into a traditional data base.

Developing techniques that allow for less precision opened a window into the untapped universe of insights. Big data, with its emphasis on comprehensive databases and messiness, helps us get closer to reality than did our dependence on small data and accuracy. Predictions based on **correlations lie at the heart of big data and are now used so frequently that we sometimes fail to appreciate their inroads.** For instance, FICO credit scores are now used to enable health providers to target clients for reminders to take their medication. **FICO is also used as a cheap proxy for people's income levels**, for the risk of illness like high blood pressure, diabetes, or depression. Lifestyle data that includes variables such as hobbies, the websites people visit, and the amount of television they watch help identify other health risks. Algorithms are being used to predict hit songs and predict mechanical or structural failures. By combining big-data collection with clever analysis one project was able to detect a deflationary swing in prices immediately after Lehman Bother filed for bankruptcy in September 2008, while those who relied on the official CPI data had to wait until November to see it.

Unlike with correlations where the math is straightforward, there is no obvious mathematical way to "prove" causality. Correlation analysis and similar non-causal methods based on hard data are superior to and faster than most intuited causal connections. It is also more useful and efficient than slow causal thinking that is epitomized by carefully controlled (and thus costly) experiments. Causality won't be discarded (especially in the legal process), but it is being knocked off its pedestal as the primary fountain of meaning. Everything is obvious once you know the answer and big data transforms how we understand and explore the world. In the age of small data, we were driven by hypotheses, which we then attempted to validate by collecting and analyzing data. **Big data** may not spell the "end of theory," but **it fundamentally transforms the way we make sense of the world.**

Datafication of a phenomenon is to put it in a quantified format so it can be tabulated and analyzed. We largely digitized text in the 1990s. More recently, as storage capacity, processing power, and bandwidth have increased, we've done it with other forms of content too, like images, video, and music.

The Medici became the most influential bankers in Europe in the 16th century largely through a superior method of data recording, the double-entry system. This along with a math textbook published by a Franciscan monk name Pacioli established the use of Arabic numerals in the West. Parallel to advances in the recording of data, ways of measuring the world—denoting time distance area, volume, and weight—continued to gain ever increasing precision. Commodore Maury's charts did so for navigation. Whenever Datafication succeeded, enormous value was created from the underlying information and tremendous insights were uncovered. Amazon understands the value of digitizing content, while Google understands the value of datafying it. While GPS is accurate to one meter, even better accuracy can be established by triangulating among cell towers or wifi routers to determine position based on signal strength, since GPS doesn't work indoors or amid tall buildings.

In the US and Britain, **drivers can buy car insurance priced according to where and when they actually drive.** It shifts the very nature of insurance from one based on pooled risk to something based on individual action. Some retailers are positioning store surveillance cameras so that they not only spot shoplifters but also track the flow of customers through the store and where they stop to look. They can use this to design the best layout for the store as well as to judge the effectiveness of marketing campaigns. Using

MOOC (massive open online courses), education programs like Udacity, Coursera and edX track the web interactions of students to see what works best pedagogically with class sizes of tens of thousands of students producing extraordinary amounts of data. They **may upend the entire education industry and pop the growing cost bubble.**

Inspectors increased their efficiency fivefold through big-data analysis **in New York City.** The most important reason for one program's success was that it dispensed with a reliance on causation in favor of correlation. Because correlations can be found far faster and cheaper than causation, they're often preferable. We will still need causal studies and controlled experiments with carefully curated data in certain cases, but for many everyday needs, knowing *what,* not *why,* is good enough.

A number of startups have looked into adapting the **social graph to use as signals for establishing credit scores.** The idea is that birds of a feather flock together: prudent people befriend like-minded types, while the profligate hang out among themselves. A Hollywood application has developed a model that looks at the rate at which new tweets were posted. With this, they are able to forecast a film's success better than other commonly used predictors. Google, Facebook, Twitter, LinkedIn, Foursquare and others sit on an enormous treasure chest of datafied information that, once analyzed, will shed light on social dynamics at all levels, from the individual to society at large. Facebook has been shrewdly patient, knowing that unveiling too many new purposes for its users' data too soon could freak them out. More of the criticism it has faced centers on what information it is capable of collecting than on what it has actually done with that data.

Aqueducts made possible the growth of cities (by providing adequate water to dense populations); the printing press facilitated the Enlightenment, and newspapers enabled the rise of the nation state. But these **infrastructures were focused on flows of water and knowledge, as were the telephone and the internet. In contrast, Datafication represents an essential enrichment in human comprehension.** We will no longer regard our world as a string of happenings that we explain as natural or social phenomena, but as a universe comprised essentially of information.

Unlike material things—the food we eat, a candle that burns--data's value does not diminish when it is used; it can be processed again and again. One person's use of it does not impede another's. And information doesn't wear out with use the way material goods do. **Data's full value is much greater than that extracted from its first use. As it moves from primary to secondary uses it becomes much more valuable over time** considered in terms of all the possible ways it can be employed in the future, not simply how it is used in the present.

When the state gathers data, it does so on behalf of its citizens, and thus it ought to provide access to society (except for security or privacy rights issues). Open government initiatives around the globe argue that **governments are only custodians of the information they collect,** and that the private sector and society will be more innovative for both civic and commercial purposes. Obama on his first day in office in January 2009 issued a memorandum ordering the head of federal agencies to release as much data as possible and on 5/9/13 released *all* federal data to the public in an executive order.

The gap between book value (hard assets) and market value of companies has been growing for decades. It has grown from about 40% of the value of publicly traded companies in the US in the mid-1980s to 75% in 2000. Increasingly intangible assess are coming to mean the data that companies hold and use. Share prices may swell for companies that have data or can collect it easily, while others in less fortunate positions may see their market valuations shrink. As accounting quandaries and liability concerns are alleviated, **it is almost certain that the value of data will show up on corporate balance sheets and emerge as a new asset class.** This resembles the difficulties of pricing financial derivatives prior to the development of the Black-Scholes equation in the 1970s, or the difficulty in valuing patents, where auctions, exchanges, private sales, licensing, and lots of litigation are slowly creating a market for knowledge. Data holders may want to opt for an arrangement that pays them a percentage of the value extracted from the data rather than a fixed fee. Most of data's value lies in its use, not its mere possession.

To be successful, you want to be complementary and scarce to something that is ubiquitous and cheap. Data is so widely available and so strategically important that the scarce thing now is the knowledge to extract wisdom from it. This puts statisticians, database managers, and machine learning people in a fantastic position. Today, in big data's early stages, the ideas and the skills seem to hold the greatest worth. But eventually most value will be in the data itself.

Statistical analyses force people to reconsider their instincts. **We are seeing the waning of subject-matter experts' influence in many areas.** In media, the content that gets created and publicized on websites like Huffington Post, Gawke, Forbes, and the Drudge Report is regularly determined by data, not just the judgment of human editors. This means that the skills needed to succeed in the workplace are changing. Expertise is like exactitude: appropriate for a small-data world where one never has enough or the right

information, and thus has to rely on intuition and experience to guide one's way. What it takes for an employee to be valuable to a company changes. **What and who you need to know changes,** and so does what you need to study to prepare for professional life. Mathematics and statistics with some programming and network science will be as foundational to the modern workplace as numeracy was a century ago and literacy before that.

Studies of the performance of **companies that excel at data-driven decision-making** have found that **productivity levels were as much as 6% higher than companies that do not.** Scale still matters, but what counts now is scale in data. "Scale without mass" allows a large virtual presence without hefty physical resources and can diffuse innovations broadly at little cost. There are scale advantages to the very large, and cost and innovation advantages to the small. The West's early lead in big data will diminish as other parts of the world adopt the technology. But the **good news for today's powerhouse firms is that big data will probably accentuate corporate strengths and weaknesses.**

Humans are primed to see the world through the lens of cause and effect. Thus, big data is under constant threat of being abused for causal purposes, rather than correlations. Steve Jobs may have continually improved the Mac laptop over years on the basis of field reports, but he used his intuition, not data, to launch the iPod, iPhone, and iPad. Robert McNamara died in 2009 at age 93, a man of intelligence (and a master of data) but not of wisdom. Correlations do not imply causation. In the big-data era **we will have to expand our understanding of justice and require that it include safeguards for human agency** much as we currently protect procedural fairness. The more we switch from holding people accountable for their acts to relying on data-driven interventions to reduce risk in society, the more we devalue the ideal of individual responsibility.

Data is to the information society what fuel was to the industrial economy. We will need measures comparable to the ones that established competition and oversight in the earlier areas of technology, such as those for the railroads in the 1800s, IBM in the 1960s, Xerox in the 1970s, AT&T in the 1980s, Microsoft in the 1990s, and Google today.

Quick correlations let us save money on plane tickets, predict flu outbreaks, and know which manholes or overcrowded buildings to inspect in a resource-constrained world. They may enable health insurance to provide coverage without a physical exam and lower the cost of reminding the sick to take their medication. If big **data teaches us** anything, it is **that just making improvements without a deeper understanding is often good enough.** Big data is a resource and a tool meant to inform, rather than explain. Our current big-data world will soon look as quaint as the 4 kilobytes of writeable memory on Apollo 11's guidance control computer does now.

[**The quantitative change has led to a qualitative one. Big data refers to things one can do at a large scale that cannot be done at a smaller one. With information, as with physics, size matters. It leads society to abandon its time-honored preference for causality,** and in many instances tap the benefits of correlation. **Social sciences have lost their monopoly on making sense of empirical social data. There is a premium on diversity within a group and in society at large. Correlations lie at the heart of big data and are now used so frequently that we sometimes fail to appreciate their inroads. FICO is also used as a cheap proxy for people's income levels. MOOC (massive open online courses) may upend the entire education industry and pop the growing cost bubble. infrastructures were focused on flows of water and knowledge, as were the telephone and the internet. In contrast, Datafication represents an essential enrichment in human comprehension. Governments are only custodians of the information they collect. It is almost certain that the value of data will show up on corporate balance sheets and emerge as a new asset class. Data is to the information society what fuel was to the industrial economy.]**