The Rise of Big Data: How It's Changing the Way We Think about the World

Kenneth Cukier and
Viktor Mayer-Schönberger

Everyone knows that the Internet has changed how businesses operate, governments function, and people live. But a new, less visible technological trend is just as transformative: "big data." Big data starts with the fact that there is a lot more information floating around these days than ever before, and it is being put to extraordinary new uses. Big data is distinct from the Internet, although the web makes it much easier to collect and share data. Big data is about more than just communication: the idea is that we can learn from a large body of information things that we could not comprehend when we used only smaller amounts.

In the third century BC, the Library of Alexandria was believed to house the sum of human knowledge. Today, there is enough information in the world to give every person alive 320 times as much of it as historians think was stored in Alexandria's entire collection—an estimated 1,200 exabytes' worth. If all this information were placed on CDs and they were stacked up, the CDs would form five separate piles that would all reach to the moon.

This explosion of data is relatively new. As recently as the year 2000, only one-quarter of all the world's stored information was digital. The rest was preserved on paper, film, and other analog media. But because the amount of digital data expands so quickly—doubling around every three years—that situation was swiftly inverted. Today, less than two percent of all stored information is nondigital.

Given this massive scale, it is tempting to understand big data solely in terms of size. But that would be misleading. Big data is also characterized by the ability to render into data many aspects of the world that have never been quantified before; call it "datafication." For example, location has been datafied, first with the invention of longitude

and latitude, and more recently with GPS satellite systems. Words are treated as data when computers mine centuries' worth of books. Even friendships and "likes" are datafied, via Facebook.

This kind of data is being put to incredible new uses with the assistance of inexpensive computer memory, powerful processors, smart algorithms, clever software, and math that borrows from basic statistics. Instead of trying to "teach" a computer how to do things, such as drive a car or translate between languages, which artificial intelligence experts have tried unsuccessfully to do for decades, the new approach is to feed enough data into a computer that it can infer the probability that, say, a traffic light is green and not red or that, in a certain context, *lumière* is a more appropriate substitute for "light" than *léger.*

Using great volumes of information in this way requires three profound changes in how we approach data. The first is to collect and use a lot of data rather than settle for small amounts or samples, as statisticians have done for well over a century. The second is to shed our preference for highly curated and pristine data and instead accept messiness: in an increasing number of situations, a bit of inaccuracy can be tolerated because the benefits of using vastly more data of variable quality outweigh the costs of using smaller amounts of very exact data. Third, in many instances, we need to give up our quest to discover the cause of things, in return for accepting correlations. With big data, instead of trying to understand precisely why an engine breaks down or why a drug's side effect disappears, researchers can instead collect and analyze massive quantities of information about such events and everything that is associated with them, looking for patterns that might help predict future occurrences. Big data helps answer what, not why, and often that's good enough.

The Internet has reshaped how humanity communicates. Big data is different: it marks a transformation in how society processes information. In time, big data might change our way of thinking about the world. As we tap ever more data to understand events and make decisions, we are likely to discover that many aspects of life are probabilistic, rather than certain.

*Approaching "***n** = *all"*

For most of history, people have worked with relatively small amounts of data because the tools for collecting, organizing, storing, and analyzing information were poor. People winnowed the information they

relied on to the barest minimum so that they could examine it more easily. This was the genius of modern-day statistics, which first came to the fore in the late nineteenth century and enabled society to understand complex realities even when few data existed. Today, the technical environment has shifted 179 degrees. There still is, and always will be, a constraint on how much data we can manage, but it is far less limiting than it used to be and will become even less so as time goes on.

The way people handled the problem of capturing information in the past was through sampling. When collecting data was costly and processing it was difficult and time-consuming, the sample was a savior. Modern sampling is based on the idea that, within a certain margin of error, one can infer something about the total population from a small subset, as long the sample is chosen at random. Hence, exit polls on election night query a randomly selected group of several hundred people to predict the voting behavior of an entire state. For straightforward questions, this process works well. But it falls apart when we want to drill down into subgroups within the sample. What if a pollster wants to know which candidate single women under 30 are most likely to vote for? How about university-educated, single Asian American women under 30? Suddenly, the random sample is largely useless, since there may be only a couple of people with those characteristics in the sample, too few to make a meaningful assessment of how the entire subpopulation will vote. But if we collect all the data—"n = all," to use the terminology of statistics—the problem disappears.

This example raises another shortcoming of using some data rather than all of it. In the past, when people collected only a few data, they often had to decide at the outset what to collect and how it would be used. Today, when we gather all the data, we do not need to know beforehand what we plan to use it for. Of course, it might not always be possible to collect all the data, but it is getting much more feasible to capture vastly more of a phenomenon than simply a sample and to aim for all of it. Big data is a matter not just of creating somewhat larger samples but of harnessing as much of the existing data as possible about what is being studied. We still need statistics; we just no longer need to rely on small samples.

There is a trade-off to make, however. When we increase the scale by orders of magnitude, we might have to give up on clean, carefully curated data and tolerate some messiness. This idea runs counter to

how people have tried to work with data for centuries. Yet the obsession with accuracy and precision is in some ways an artifact of an informationconstrained environment. When there were not that many data around, researchers had to make sure that the figures they bothered to collect were as exact as possible. Tapping vastly more data means that we can now allow some inaccuracies to slip in (provided the data set is not completely incorrect), in return for benefiting from the insights that a massive body of data provides.

Consider language translation. It might seem obvious that computers would translate well, since they can store lots of information and retrieve it quickly. But if one were to simply substitute words from a French-English dictionary, the translation would be atrocious. Language is complex. A breakthrough came in the 1990s, when IBM delved into statistical machine translation. It fed Canadian parliamentary transcripts in both French and English into a computer and programmed it to infer which word in one language is the best alternative for another. This process changed the task of translation into a giant problem of probability and math. But after this initial improvement, progress stalled.

Then Google barged in. Instead of using a relatively small number of high-quality translations, the search giant harnessed more data, but from the less orderly Internet—"data in the wild," so to speak. Google inhaled translations from corporate websites, documents in every language from the European Union, even translations from its giant book-scanning project. Instead of millions of pages of texts, Google analyzed billions. The result is that its translations are quite good—better than IBM's were—and cover 65 languages. Large amounts of messy data trumped small amounts of cleaner data.

From Causation to Correlation

These two shifts in how we think about data—from some to all and from clean to messy—give rise to a third change: from causation to correlation. This change represents a move away from always trying to understand the deeper reasons behind how the world works to simply learning about an association among phenomena and using that to get things done.

Of course, knowing the causes behind things is desirable. The problem is that causes are often extremely hard to figure out, and many

times, when we think we have identified them, it is nothing more than a self-congratulatory illusion. Behavioral economics has shown that humans are conditioned to see causes even where none exist. So we need to be particularly on guard to prevent our cognitive biases from deluding us; sometimes, we just have to let the data speak.

Take UPS, the delivery company. It places sensors on vehicle parts to identify certain heat or vibrational patterns that in the past have been associated with failures in those parts. In this way, the company can predict a breakdown before it happens and replace the part when it is convenient, instead of on the side of the road. The data do not reveal the exact relationship between the heat or the vibrational patterns and the part's failure. They do not tell UPS why the part is in trouble. But they reveal enough for the company to know what to do in the near term and guide its investigation into any underlying problem that might exist with the part in question or with the vehicle.

A similar approach is being used to treat breakdowns of the human machine. Researchers in Canada are developing a big-data approach to spot infections in premature babies before overt symptoms appear. By converting 16 vital signs, including heartbeat, blood pressure, respiration, and blood-oxygen levels, into an information flow of more than 1,000 data points per second, they have been able to find correlations between very minor changes and more serious problems. Eventually, this technique will enable doctors to act earlier to save lives. Over time, recording these observations might also allow doctors to understand what actually causes such problems. But when a newborn's health is at risk, simply knowing that something is likely to occur can be far more important than understanding exactly why.

Medicine provides another good example of why, with big data, seeing correlations can be enormously valuable, even when the underlying causes remain obscure. In February 2009, Google created a stir in healthcare circles. Researchers at the company published a paper in *Nature* that showed how it was possible to track outbreaks of the seasonal flu using nothing more than the archived records of Google searches. Google handles more than a billion searches in the United States every day and stores them all. The company took the 50 million most commonly searched terms between 2003 and 2008 and compared them against historical influenza data from the Centers for Disease Control and Prevention. The idea was to discover whether the incidence of certain searches coincided

with outbreaks of the flu—in other words, to see whether an increase in the frequency of certain Google searches conducted in a particular geographic area correlated with the CDC's data on outbreaks of flu there. The CDC tracks actual patient visits to hospitals and clinics across the country, but the information it releases suffers from a reporting lag of a week or two—an eternity in the case of a pandemic. Google's system, by contrast, would work in near-real time.

Google did not presume to know which queries would prove to be the best indicators. Instead, it ran all the terms through an algorithm that ranked how well they correlated with flu outbreaks. Then the system tried combining the terms to see if that improved the model. Finally, after running nearly half a billion calculations against the data, Google identified 45 terms—words such as "headache" and "runny nose"—that had a strong correlation with the CDC's data on flu outbreaks. All 45 terms related in some way to influenza. But with a billion searches a day, it would have been impossible for a person to guess which ones might work best and test only those.

Moreover, the data were imperfect. Since the data were never intended to be used in this way, misspellings and incomplete phrases were common. But the sheer size of the data set more than compensated for its messiness. The result, of course, was simply a correlation. It said nothing about the reasons why someone performed any particular search. Was it because the person felt ill, or heard sneezing in the next cubicle, or felt anxious after reading the news? Google's system doesn't know, and it doesn't care. Indeed, last December, it seems that Google's system may have overestimated the number of flu cases in the United States. This serves as a reminder that predictions are only probabilities and are not always correct, especially when the basis for the prediction—Internet searches—is in a constant state of change and vulnerable to outside influences, such as media reports. Still, big data can hint at the general direction of an ongoing development, and Google's system did just that.

Back-End Operations

Many technologists believe that big data traces its lineage back to the digital revolution of the 1980s, when advances in microprocessors and computer memory made it possible to analyze and store ever more

information. That is only superficially the case. Computers and the Internet certainly aid big data by lowering the cost of collecting, storing, processing, and sharing information. But at its heart, big data is only the latest step in humanity's quest to understand and quantify the world. To appreciate how this is the case, it helps to take a quick look behind us.

Appreciating people's posteriors is the art and science of Shigeomi Koshimizu, a professor at the Advanced Institute of Industrial Technology in Tokyo. Few would think that the way a person sits constitutes information, but it can. When a person is seated, the contours of the body, its posture, and its weight distribution can all be quantified and tabulated. Koshimizu and his team of engineers convert backsides into data by measuring the pressure they exert at 360 different points with sensors placed in a car seat and by indexing each point on a scale of zero to 256. The result is a digital code that is unique to each individual. In a trial, the system was able to distinguish among a handful of people with 98 percent accuracy.

The research is not asinine. Koshimizu's plan is to adapt the technology as an antitheft system for cars. A vehicle equipped with it could recognize when someone other than an approved driver sat down behind the wheel and could demand a password to allow the car to function. Transforming sitting positions into data creates a viable service and a potentially lucrative business. And its usefulness may go far beyond deterring auto theft. For instance, the aggregated data might reveal clues about a relationship between drivers' posture and road safety, such as telltale shifts in position before accidents. The system might also be able to sense when a driver slumps slightly from fatigue and send an alert or automatically apply the brakes.

Koshimizu took something that had never been treated as data—or even imagined to have an informational quality—and transformed it into a numerically quantified format. There is no good term yet for this sort of transformation, but "datafication" seems apt. Datafication is not the same as digitization, which takes analog content—books, films, photographs—and converts it into digital information, a sequence of ones and zeros that computers can read. Datafication is a far broader activity: taking all aspects of life and turning them into data. Google's augmented-reality glasses "datafy" the gaze. Twitter datafies stray thoughts. LinkedIn datafies professional networks.

Once we datafy things, we can transform their purpose and turn the information into new forms of value. For example, IBM was granted a U.S. patent in 2012 for "securing premises using surfacebased computing technology"—a technical way of describing a touchsensitive floor covering, somewhat like a giant smartphone screen. Datafying the floor can open up all kinds of possibilities. The floor could be able to identify the objects on it, so that it might know to turn on lights in a room or open doors when a person entered. Moreover, it might identify individuals by their weight or by the way they stand and walk. It could tell if someone fell and did not get back up, an important feature for the elderly. Retailers could track the flow of customers through their stores. Once it becomes possible to turn activities of this kind into data that can be stored and analyzed, we can learn more about the world—things we could never know before because we could not measure them easily and cheaply.

Big Data in the Big Apple

Big data will have implications far beyond medicine and consumer goods: it will profoundly change how governments work and alter the nature of politics. When it comes to generating economic growth, providing public services, or fighting wars, those who can harness big data effectively will enjoy a significant edge over others. So far, the most exciting work is happening at the municipal level, where it is easier to access data and to experiment with the information. In an effort spearheaded by New York City Mayor Michael Bloomberg (who made a fortune in the data business), the city is using big data to improve public services and lower costs. One example is a new fire-prevention strategy.

Illegally subdivided buildings are far more likely than other buildings to go up in flames. The city gets 25,000 complaints about overcrowded buildings a year, but it has only 200 inspectors to respond. A small team of analytics specialists in the mayor's office reckoned that big data could help resolve this imbalance between needs and resources. The team created a database of all 900,000 buildings in the city and augmented it with troves of data collected by 19 city agencies: records of tax liens, anomalies in utility usage, service cuts, missed payments, ambulance visits, local crime rates, rodent complaints, and more. Then they compared this database to records of building fires from the past five years,

ranked by severity, hoping to uncover correlations. Not surprisingly, among the predictors of a fire were the type of building and the year it was built. Less expected, however, was the finding that buildings that obtained permits for exterior brickwork correlated with lower risks of severe fire.

Using all this data allowed the team to create a system that could help them determine which overcrowding complaints needed urgent attention. None of the buildings' characteristics they recorded caused fires; rather, they correlated with an increased or decreased risk of fire. That knowledge has proved immensely valuable: in the past, building inspectors issued vacate orders in 13 percent of their visits; using the new method, that figure rose to 70 percent—a huge efficiency gain.

Of course, insurance companies have long used similar methods to estimate fire risks, but they mainly rely on only a handful of attributes and usually ones that intuitively correspond with fires. By contrast, New York City's big-data approach was able to examine many more variables, including ones that would not at first seem to have any relation to fire risk. And the city's model was cheaper and faster, since it made use of existing data. Most important, the big-data predictions are probably more on target, too.

Big data is also helping increase the transparency of democratic governance. A movement has grown up around the idea of "open data," which goes beyond the freedom-of-information laws that are now commonplace in developed democracies. Supporters call on governments to make the vast amounts of innocuous data that they hold easily available to the public. The United States has been at the forefront, with its data.gov website, and many other countries have followed.

At the same time as governments promote the use of big data, they also need to protect citizens against unhealthy market dominance. Companies such as Google, Amazon, and Facebook—as well as lesser known "data brokers," such as Acxiom and Experian—are amassing vast amounts of information on everyone and everything. Antitrust laws protect against the monopolization of markets for goods and services such as software or media outlets because the sizes of the markets for those goods are relatively easy to estimate. But how should governments apply antitrust rules to big data, a market that is hard to define and that is constantly changing form? Meanwhile, privacy will become an even bigger worry, since more data will almost certainly lead to

more compromised private information, a downside of big data that current technologies and laws seem unlikely to prevent.

Regulations governing big data might even emerge as a battleground among countries. European governments are already scrutinizing Google over a raft of antitrust and privacy concerns in a scenario reminiscent of the antitrust enforcement actions the European Commission took against Microsoft beginning a decade ago. Facebook might become a target for similar actions all over the world because it holds so much data about individuals. Diplomats should brace for fights over whether to treat information flows as similar to free trade: in the future, when China censors Internet searches, it might face complaints not only about unjustly muzzling speech but also about unfairly restraining commerce.

Big Data or Big Brother?

States will need to help protect their citizens and their markets from new vulnerabilities caused by big data. But there is another potential dark side: big data could become Big Brother. In all countries, but particularly in nondemocratic ones, big data exacerbates the existing asymmetry of power between the state and the people.

The asymmetry could well become so great that it leads to big-data authoritarianism, a possibility vividly imagined in science fiction movies such as *Minority Report*. That 2002 film took place in a nearfuture dystopia in which the character played by Tom Cruise headed a "Precrime" police unit that relied on clairvoyants whose visions identified people who were about to commit crimes. The plot revolves around the system's obvious potential for error and, worse yet, its denial of free will.

Although the idea of identifying potential wrongdoers before they have committed a crime seems fanciful, big data has allowed some authorities to take it seriously. In 2007, the Department of Homeland Security launched a research project called FAST (Future Attribute Screening Technology), aimed at identifying potential terrorists by analyzing data about individuals' vital signs, body language, and other physiological patterns. Police forces in many cities, including Los Angeles, Memphis, Richmond, and Santa Cruz, have adopted "predictive policing" software, which analyzes data on previous crimes to identify where and when the next ones might be committed.

For the moment, these systems do not identify specific individuals as suspects. But that is the direction in which things seem to be heading. Perhaps such systems would identify which young people are most likely to shoplift. There might be decent reasons to get so specific, especially when it comes to preventing negative social outcomes other than crime. For example, if social workers could tell with 95 percent accuracy which teenage girls would get pregnant or which high school boys would drop out of school, wouldn't they be remiss if they did not step in to help? It sounds tempting. Prevention is better than punishment, after all. But even an intervention that did not admonish and instead provided assistance could be construed as a penalty—at the very least, one might be stigmatized in the eyes of others. In this case, the state's actions would take the form of a penalty before any act were committed, obliterating the sanctity of free will.

Another worry is what could happen when governments put too much trust in the power of data. In his 1999 book, *Seeing Like a State,* the anthropologist James Scott documented the ways in which governments, in their zeal for quantification and data collection, sometimes end up making people's lives miserable. They use maps to determine how to reorganize communities without first learning anything about the people who live there. They use long tables of data about harvests to decide to collectivize agriculture without knowing a whit about farming. They take all the imperfect, organic ways in which people have interacted over time and bend them to their needs, sometimes just to satisfy a desire for quantifiable order.

This misplaced trust in data can come back to bite. Organizations can be beguiled by data's false charms and endow more meaning to the numbers than they deserve. That is one of the lessons of the Vietnam War. U.S. Secretary of Defense Robert McNamara became obsessed with using statistics as a way to measure the war's progress. He and his colleagues fixated on the number of enemy fighters killed. Relied on by commanders and published daily in newspapers, the body count became the data point that defined an era. To the war's supporters, it was proof of progress; to critics, it was evidence of the war's immorality. Yet the statistics revealed very little about the complex reality of the conflict. The figures were frequently inaccurate and were of little value as a way to measure success. Although it is important to learn from data to improve lives, common sense must be permitted to override the spreadsheets.

Human Touch

Big data is poised to reshape the way we live, work, and think. A worldview built on the importance of causation is being challenged by a preponderance of correlations. The possession of knowledge, which once meant an understanding of the past, is coming to mean an ability to predict the future. The challenges posed by big data will not be easy to resolve. Rather, they are simply the next step in the timeless debate over how to best understand the world.

Still, big data will become integral to addressing many of the world's pressing problems. Tackling climate change will require analyzing pollution data to understand where best to focus efforts and find ways to mitigate problems. The sensors being placed all over the world, including those embedded in smartphones, provide a wealth of data that will allow climatologists to more accurately model global warming. Meanwhile, improving and lowering the cost of health care, especially for the world's poor, will make it necessary to automate some tasks that currently require human judgment but could be done by a computer, such as examining biopsies for cancerous cells or detecting infections before symptoms fully emerge.

Ultimately, big data marks the moment when the "information society" finally fulfills the promise implied by its name. The data take center stage. All those digital bits that have been gathered can now be harnessed in novel ways to serve new purposes and unlock new forms of value. But this situation requires a new way of thinking and will challenge institutions and identities. In a world where data shape decisions more and more, what purpose will remain for people, or for intuition, or for going against the facts? If everyone appeals to the data and harnesses big-data tools, perhaps what will become the central point of differentiation is unpredictability: the human element of instinct, risk taking, accidents, and even error. If so, then there will be a special need to carve out a place for the human: to reserve space for intuition, common sense, and serendipity to ensure that they are not crowded out by data and machine-made answers.

This possibility has important implications for the notion of progress in society. Big data enables us to experiment faster and explore more leads. These advantages should produce more innovation. But at times, the spark of invention becomes what the data do not say. That

is something that no amount of data can ever confirm or corroborate, since it has yet to exist. If Henry Ford had queried big-data algorithms to discover what his customers wanted, they would have come back with "a faster horse," to recast his famous line. In a world of big data, it is the most human traits that will need to be fostered—creativity, intuition, and intellectual ambition—since human ingenuity is the source of progress.

Big data is a resource and a tool. It is meant to inform, rather than explain; it points toward understanding, but it can still lead to misunderstanding, depending on how well it is wielded. And however dazzling the power of big data appears, its seductive glimmer must never blind us to its inherent imperfections. Rather, we must adopt this technology with an appreciation not just of its power but also of its limitations.