THE FUTURE OF DATA-DRIVEN INNOVATION

U.S. CHAMBER OF COMMERCE FOUNDATION

U.S. CHAMBER OF COMMERCE FOUNDATION

The U.S. Chamber of Commerce Foundation (USCCF) is a 501(c)(3) nonprofit affiliate of the U.S. Chamber of Commerce dedicated to strengthening America's long-term competitiveness and educating the public on how our free enterprise system improves society and the economy.

THE FUTURE OF DATA-DRIVEN INNOVATION

Presented By
The U. S. Chamber of Commerce Foundation

TABLE OF CONTENTS

A LETTER FROM

USCCF PRESIDENT JOHN R. MCKERNAN, JR.

U.S. CHAMBER OF COMMERCE FOUNDATION

John R. McKernan, Jr.
President

1615 H Street, NW
Washington, DC 20062-2000
202-463-5946
JMcKernan@uschamber.com
foundation.uschamber.com

October 7, 2014

Dear Colleague:

It is the U.S. Chamber of Commerce Foundation's mission to explore America's long-term competitiveness and educate the public on how the free enterprise system improves society and the economy. Part of that mission charges us to explore the emerging issues that are changing the ways we do business in this country and around the world. There is arguably no bigger emerging issue in today's business environment than the rise of data-driven innovation.

From collection and access to technology and privacy, the implications surrounding the issue of data impact every continent, industry, community, and citizen in some shape or fashion. No one is left untouched by this environment as data permeates everything around us. That is both revolutionary and exciting but also of concern to many, and it is a reality that deserves both our attention and an informed discussion.

Regardless of what form it takes, data tells a story. It can identify cost savings and efficiencies, new connections and opportunities, and an improved understanding of the past to shape a better future. It also provides the details necessary to allow us to make more informed decisions about the next step we want to take. These are the benefits of the unfolding data revolution and the good it offers to us all. Those good things, however, spur dialogue and debate across a range of areas, and it is why we in the Foundation took a focused look at the issues and innovations that are happening in data today.

We started this effort early in 2014 by talking with private sector leaders of enterprises small and large, with government officials at all levels, and with various educators, analysts, and other experts from around the country. From those numerous discussions, events, and programs, we identified several notable experts who offered to share their own insights on these issues and how they will shape the future of data-driven innovation and the economy around it. This report shares the thoughts and insights of these various practitioners. While none of them offers the absolute final word on the data-driven economy, the future of competitiveness, or the policies that would enable more innovation to happen, they do help us better inform the conversation that needs to be had on these issues.

At the Chamber Foundation, we believe that information and discussion is the only way to better understand the emerging issues that data and all of its offerings provide to our future. If the United States is to continue to lead in these areas, it has to be through an active and informed conversation driven by facts, details, and real-world experiences. Going forward, the Foundation will continue to share its insights in these and other areas. By sharing ideas with one another, we know we have a data-driven future for good that will change lives around the world for the better.

Sincerely,

John R. McKernan, Jr.
President, U.S. Chamber of Commerce Foundation

INTRODUCTION

DEFINING THE DATA MOVEMENT

BY JUSTIN HIENZ

Today, there is a rapidly growing capacity to collect, store and analyze massive amounts of data, far more than an individual mind could process on its own. This enormous volume of information has been called Big Data, a term that is widely used, sometimes to the point of cliché. Yet, while the term can be trite, the dramatic potential in exploring large datasets for new insights, trends, and correlations is anything but.

The data movement is a force for good. It is fodder for research and a catalyst for innovation. It is the bedrock of informed decision-making and better business and the key to unlocking more efficient, effective government and other services. It unleashes economic growth, competition, profitability, and other breakthrough discoveries. And it is at once a product of an ever-more technologically sophisticated world and a tool to advance, enhance, and shape all of its domains going forward. This widespread emergence and use of Big Data is revolutionary, and history will record the early 21st century as the beginning of a data revolution that defined a century.

There is no shortage of examples of data-driven decision making and innovation. Less common, however, is scholarship that looks at myriad examples to extrapolate the ideas, themes and potential that define the data movement and the changes it will bring. This report begins to fill that knowledge gap, with leading scholars and practitioners looking to the horizon to describe the data-driven future.

The world is but a few steps down the data road. In time, the very notion of "Big Data" will fade as data-driven decision making becomes a ubiquitous and unquestioned piece of everyday life. Yet, the way we understand and embrace the data movement now will shape how it impacts all of our futures. This report informs the ongoing discussion to reveal how data impacts our lives, economies, societies, the choices we make, and, inevitably, changes everything for the better.

The Rise of Big Data

Because this is only the beginning of the data-driven movement, the terms, definitions, and ideas associated with Big Data are still evolving. This is a subject area that is as dynamic as it is amorphous. One can no more wrap their arms around a tidal wave than

"THE AMOUNT OF DATA GENERATED IN TWO DAYS IS AS MUCH AS ALL DATA GENERATED IN HUMAN HISTORY BEFORE 2003."

they can nail down precisely all that is grouped under the Big Data epithet. That said, there are some clear properties of the data landscape.

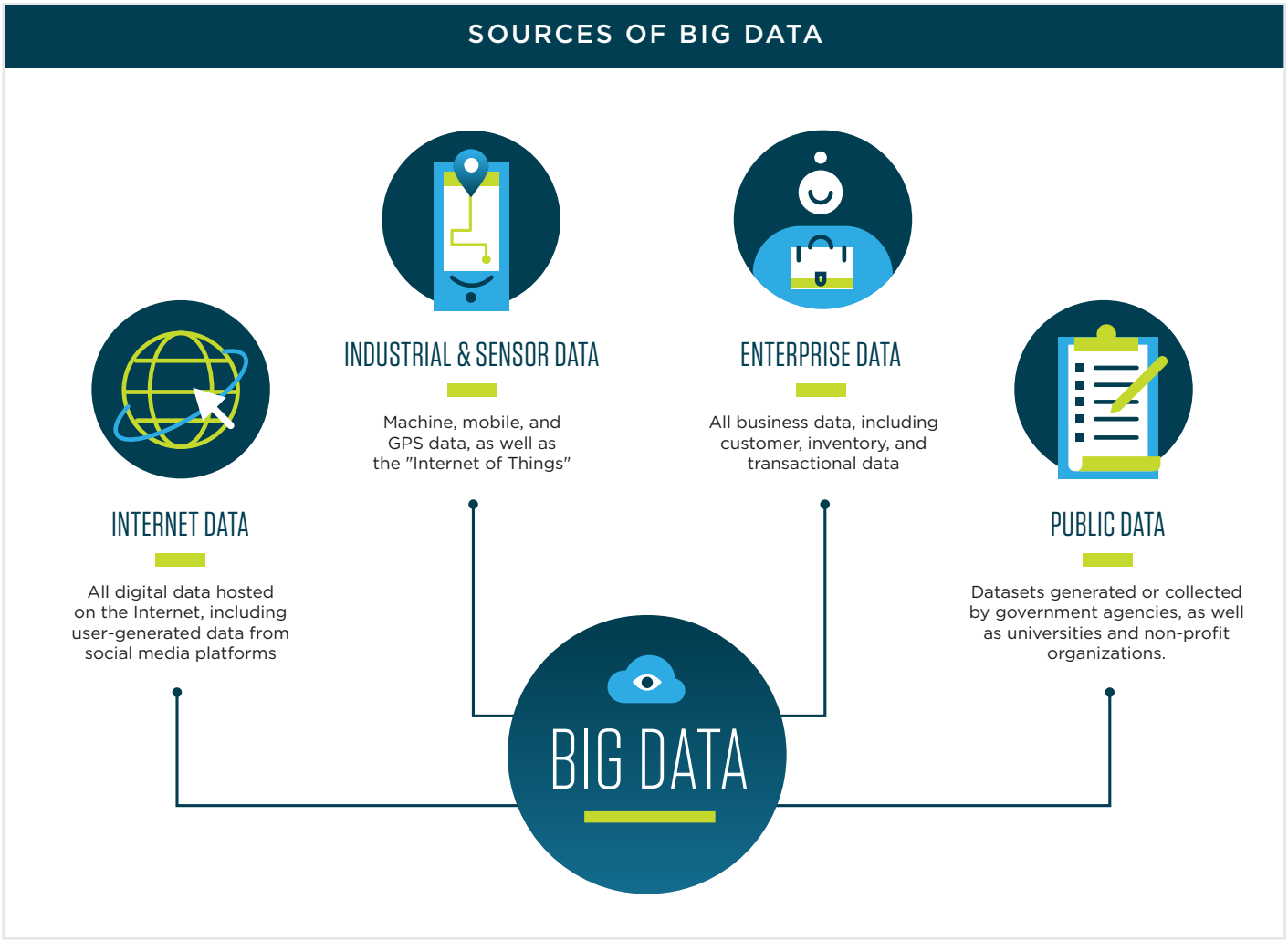
Most definitions of Big Data draw from Doug Laney's often-cited "three Vs," each of which describe a component quality of Big Data: volume (the amount of data); velocity (the speed at which data is created); and variety (the types of data).[1] Of late, Big Data definitions have come to include a fourth V: veracity. As shown throughout this report, data accuracy is as important to realizing value as the size, type, and generation of information.

Big Data is so voluminous and generated at such a rapid pace that it cannot be effectively gathered, searched, or even understood by the human mind alone. As such, technology is the medium through which the data movement thrives. Data's volume and variety owes to the growing number of sensors and connected devices that permeate every aspect of industrialized society (the so-called "Internet of Things"). In addition, the digitization of commercial transactions, medical records, online social communication, and other information also contributes to the amount of data. And then there are datasets from governments, research groups, and international organizations, all of which create and consume data in their activities.

Collectively, in 2012, these and other sources generated 2.5 quintillion bytes of data every day.[2] To put that in perspective, as Google CEO Eric Schmidt said at a 2010 Techonomy conference, the amount of data generated in two days is as much as all data generated in human history before 2003.[3] Thousands of years of civilization, millions of books, every piece of information from the ancient Library of Alexandria to the modern Library of Congress, all of that data together is but a miniscule drop in the proverbial bucket.

Given the volume, growing alongside data generation has been the capacity to hold and access it. Ever-more ubiquitous Internet makes it possible to easily share information without regard to geographic distance or observed borders, a capability unique to our modern age that allows all professionals, businesses, and organizations to share, collaborate, and advance knowledge like never before. These online advantages, as well as advances in computing power, facilitated innovative approaches to storing the exponentially growing volume of data. This has made it practical to keep and analyze entire datasets (rather than just down-sampled portions), dramatically expanding the power and promise of Big Data.

Even as the world has made great technological advances in collecting and storing data, once in hand, making sense of all this information is a challenge unto itself. Data scientists are in great demand, with many businesses and organizations creating data-specific positions and hiring data experts who can use complex algorithms and computer programs to find correlations and trends within troves of information. Indeed, while Big Data and technology are natural bedfellows, the human element is not obsolete. Leslie Bradshaw writes in Chapter 2 that until computers can "think" creatively and contextually, the human brain will remain a critical component in turning raw data into actionable insight. Indeed, data does not



replace human thinking; it enhances it. The brain is the vehicle for innovation, and data is its fuel.

A Diverse Data Landscape

Big Data is a broad, inclusive term. It may refer to spreadsheets packed with numbers, but it also includes product reviews online, the results of procedures in medical files, or the granular data that accompanies all online activity (called metadata). Shopping histories, biological responses to new pharmaceuticals, weather and agriculture trends, manufacturing plant efficiency—these and many other kinds of information make up the complex, varied data landscape.

Data is an asset. As such, much of the data generated every day is proprietary. An online retailer owns the data listing its customers' purchases, and a pharmaceutical company owns data from testing its products. This is appropriate, since businesses bear costs to generate, store and analyze data and then enjoy the innovative fruits that grow out of it.

Yet, there is also great value in providing data freely to any interested party, a philosophy called Open Data. Whereas Big Data is generally defined by its size, Open Data (which can be "big") is defined by its use. There are two core components of Open Data: data is publicly available, licensed in a way that allows for reuse; and data is relatively easy to use (e.g., accessible, digitized, etc.). Free or low-cost data offered in a widely useable format unleashes enormous potential. It exposes data to more minds, interests, goals, and perspectives. If two heads are better than one, how much better are dozens, hundreds, or even thousands of minds digging through data, looking for valuable correlations and insights?

Joel Gurin writes in Chapter 5 that there are four kinds of Open Data driving innovation: scientific; social; personal; and governmental. From genomics to astronomy, researchers are taking a collaborative approach to working with scientific data. Meanwhile, businesses and other

organizations are seeking insight via social data (e.g., blogs, company reviews, social media posts), which can reveal consumer opinions on products, services, and brands. And new digital applications are giving citizens access to their own personal data, yielding more informed consumers.

Perhaps the most common (and most robust) Open Data, however, comes from the public sector. For example, in June 2014, the U.S. Food and Drug Administration (FDA) launched OpenFDA, a portal through which anyone can access publicly available FDA data. This initiative is designed, in the words of FDA Chief Health Informatics Officer Taha Dr. Kass-Hout, to "serve as a pilot for how FDA can interact internally and with external stakeholders, spur innovation, and develop or use novel applications securely and efficiently."[4]

A subset of the Open Data concept is what McKinsey calls MyData.[5] This is the consumer-empowering idea of sharing data that has been collected about an individual with that individual. MyData fosters transparency, informs consumers, and has myriad impacts on commerce and cost of living. McKinsey offers the example of utility companies comparing individual and aggregated statistics to show consumers how their energy use stacks up against their neighbors. Personal healthcare data is another example where sharing MyData with an individual can yield greater understanding of one's wellness and habits.

Yet, despite the evident benefits, it is when discussing this kind of data that concerns are sometimes voiced over the potential for personal information to be used to the detriment of the individual. In the public data discussion, this anxiety is often reduced to an overly simplistic cry for "privacy." When the data-privacy nexus is approached from a scholarly, objective viewpoint, however, it becomes clear that data does not present an inherent threat to privacy. Rather, the relationship between the individual and the public and private sector organizations that hold data about them is best viewed as a form of trusteeship. When the trustee (the organization) fails to uphold obligations for securing and fairly using personal data, they commit what Benjamin Wittes and Wells Bennett call "databuse" (see Chapter 6). This potential misuse of data is most deserving of discussion, rather than a fear-driven debate over vague visions of privacy.

Data-Driven Innovation

Data is a resource, much like water or energy, and like any resource, data does nothing on its own. Rather, it is world-changing in how it is employed in human decision making. Without data, decisions are guesses; with it, decisions are targeted, strategic, and informed. These lead to better business, better government, and better solutions to address the world's woes and raise its welfare.

Data has attracted the excitement and attention it has because of the massive potential in its application. Data-driven innovation, as Dr. Joseph Kennedy describes in Chapter 1, has enormous economic value, with Big Data product and service sales exceeding $18 billion in 2013, expected to reach $50 billion by 2017.[6] This value comes in the form of: new goods and services; optimized production processes and supply chains; targeted marketing; improved organizational management; faster research and development; and much more. It could include companies developing consumer products based on customer surveys, energy producers using geological studies to find oil, or

"DATA IS A RESOURCE, MUCH LIKE WATER OR ENERGY, AND LIKE ANY RESOURCE, DATA DOES NOTHING ON ITS OWN. RATHER, IT IS WORLD-CHANGING IN HOW IT IS EMPLOYED IN HUMAN DECISION MAKING."

financial firms using corporate data to advise investors.

As well as new products and services, Big Data also yields value through increased competitiveness. As John Raidt writes in Chapter 3, the United States enjoys numerous attributes that will allow the country to take the fullest advantage of the data movement, more so than any other country. The United States' longstanding technological leadership, its free market system, its research and development infrastructure, its rule of law, and a host of other national qualities make the country the most fertile for data-driven innovation and all the economic, societal, and competitive benefits that come with it.

Accessing all this value, however, depends in part on the policies guiding data gathering, usage and transmission. Matthew Harding writes in Chapter 4 that the United States (and the world) requires public policies that foster innovation and growth while protecting individual freedom and restricting potential "databuse." Many of these policies already exist in an effective form. In any case, the policies we set and uphold today will in part define how data is used in the future. There is an important role for public policy in this emerging phenomenon, but to extract the most value from data, policies must be developed carefully and in collaboration with private sector partners.

The data discussion today is less about where it started and what it is and more about where it's going. With the above definitions and descriptions commonly agreed across industries, this report takes the next step, gathering ideas and examples to describe all the ways data is contributing to a stronger economy, improved business and government, and a steady flow of world-changing innovations. The Big Data landscape is an exciting place, and the chapters that follow offer a window into how precisely data is changing everything around us—for the better.

ENDNOTES

1 Doug Laney, "3-D Data Management: Controlling Data Volume, Velocity and Variety," *Gartner, Inc.*, 6 Feb. 2001.

2 IBM, "Bringing Big Data to the Enterprise," <http://www-01.ibm.com/software/sg/data/bigdata/> (16 Aug. 2014).

3 Marshall Kirkpatrick, "Google, Privacy and the New Explosion of Data," *Techonomy*, 4 Aug. 2010.

4 Taha Kass-Hout, "OpenFDA: Innovative Initiative Opens Door to Wealth of FDA's Publicly Available Data," *openFDA*, 2 June 2014.

5 James Manyika et al., "Open Data: Unlocking Innovation and Performance with Liquid Information," *McKinsey Global Institute*, Oct. 2013.

6 Jeff Kelly, "Big Data Vendor Revenue and Market Forecast," *Wikibon*, 12 Feb. 2014.

ABOUT THE AUTHOR

Justin Hienz is the owner of Cogent Writing, LLC, a strategic content company working with businesses, organizations and thought leaders to advance their goals through writing. Hienz's self-authored and ghostwritten work has appeared in *Foreign Policy* magazine, *The Chicago Tribune, The Newark Star-Ledger, CNBC, Roll Call, Homeland Security Today,* and numerous trade and company-owned blogs and websites. He holds master's degrees in journalism and religion from the University of Missouri.

CHAPTER
1

ABOUT THE AUTHOR

Dr. Joseph Kennedy is president of Kennedy Research, LLC and a senior fellow at the Information Technology and Innovation Foundation. His main areas of experience include the impact of technology on society, macroeconomic policy, and finance. Previous positions include general counsel of the U.S. Senate Permanent Subcommittee on Investigations and chief economist for the Department of Commerce.

THE DATA-DRIVEN ECONOMY

BY DR. JOSEPH KENNEDY

Key Takeaways

The data-driven economy is capable of generating a large amount of economic value. McKinsey estimates that better use of data could increase world income by $3 trillion each year in seven industries alone.

The economic value of data is significantly increased if it is shared. Policies should strongly encourage the movement of data between functions and institutions while ensuring that ownership, security, and privacy concerns are met.

As with other resources, data will create value only when it is used to enhance goods and services that meet the needs of customers. This will require innovation on the part of government and companies.

Knowledge has always played a crucial role in economic activity and higher living standards. Human civilization is closely linked with the ability to transmit and record information. Similarly, scientific advancements depend on the increased use of objective data (rather than subjective dogma) as the best guide to understanding the world. The close link between data and our ability to intelligently shape our lives remains strong. Fortunately, our capacities are about to radically improve as new technologies and greater access to more and better data makes it possible to understand, control, and change much more of the world. This will have significant effects on the economy and living standards.

For millennia, massive amounts of oil and natural gas lay trapped within shale deposits underneath the United States, making no contribution to economic growth. Within the last decade, new technology has allowed us to exploit these previously inaccessible resources, even as some voices are decrying risks that come with taking advantage of this energy resource.[1]

Big Data, as a resource, presents similar opportunities—and corresponding, less-than-rational concerns over potential consequences of putting data to work. These concerns are misplaced. More than any other newly tapped resource, Big Data has the potential to deliver large economic gains.

The benefits of this resource (but not its costs) increase rapidly as data is shared. The central challenge for public and private sector leaders is to apply this resource to the large variety of problems that now confront us while minimizing the relatively manageable risks associated with the greater availability of data. Just as the builders of the first oil well in Titusville, Pa., could not have envisioned the combustion engine and airplanes, we cannot foresee all the uses of cheap, abundant data. Yet, we can expect a wide variety of new products and processes that add economic value. Some of these improvements will increase traditional measures of economic growth. Others will have the primary effect of reducing costs and increasing consumer surplus. In every case, however, the advent of the Big Data era is bringing with it enormous economic potential.

"THE ECONOMIC IMPACT OF BIG DATA WILL TAKE A NUMBER OF FORMS. A RECENT MCKINSEY REPORT ESTIMATES THAT IMPROVED USE OF DATA COULD GENERATE $3 TRILLION IN ADDITIONAL VALUE EACH YEAR IN SEVEN INDUSTRIES."

Estimating Big Data's Economic Impact

While Big Data is having a significant impact on the economy, that impact is difficult to measure. One reason is that the domestic use and international exchanges of data do not always show up in economic statistics.[2] If access to large amounts of data is used to build a new business that sells consumer data to advertisers, the fees advertisers pay for the information will be counted in national income. But if better data allows hotels to meet the individual preferences of their guests without charging them more, all of the benefit will be captured as consumer surplus. The customer is better off, but because the value of economic transactions remains the same, national income is unchanged.

If improved visibility into its supply chains lets a retailer cut its prices in half, national income and perhaps employment would actually fall, at least until customers spent the savings on other items. As discussed in Chapter 4 of this report, Big Data can also increase competitiveness. Companies that increase value to customers without increasing cost will likely gain market share from their rivals. If U.S. companies take market share from foreign businesses, national income would rise.

The value of Big Data is closely tied to the growing Internet of Things—the integration of sensors and transmission capability into a wide variety of objects.[3] It therefore benefits from continued progress in making sensors, transmission capacity, storage, and processing power significantly cheaper and better over time. This progress has led to a significant increase in data generation and capture from a number of diverse sources, including financial transactions, social media, traffic patterns, medical treatments, and environmental conditions.

As data becomes more accessible, it will affect the economy in a number of ways, all of which can be loosely encompassed as being part of the data-driven economy. These impacts include:[4]

Generating new goods and services, such as GM's OnStar or custom-tailored clothing, in which information is either the product itself or it contributes significantly to the quality of another product.

Optimizing production processes and supply chains, such as what Walmart has done with its stores.

Targeted marketing, including the integration of customer feedback into product design.

Improved organizational management often in the form of using data to make better decisions.

Faster research and development, which shortens the trial and error process of innovation.

The economic impact of Big Data will take a number of forms. A recent McKinsey report estimates that improved use of data could generate $3 trillion in additional value each year in seven industries.[5] Of this, $1.3 trillion would benefit the United States. McKinsey also estimates that more than half of this value will go directly to consumers in the form of things like shorter wait times in traffic, improved ability to comparison shop, and better matching between schools and students. The rest will go to companies that either create new products centered around the use of data or use data to gain an edge over their competitors. Walmart, GM, and other companies are already using Big Data to offer new products, improve their margins, and take market share from their rivals. Walmart's use of Big Data to streamline and improve its supply chain, for example, has led to a 16% increase in revenue over the last four years.[6]

A study led by Erik Brynjolfsson at the Massachusetts Institute of Technology found that firms that adopt data-driven decision making achieve output and productivity that is 5% to 6% higher than what would be expected given their other investments and use of information technology.[7] This advantage also applies to other business measures, including asset utilization, return on equity, and market value.

Access to better data may also improve the economic climate within which businesses operate. At present, macroeconomic policy is hobbled by the limitations of official government data. Accurate data is often associated with long lags, making it difficult for policymakers to know where the economy is, let alone where it is going. Although the most recent recession officially began in December 2007, the Bureau of Economic Analysis reported as late as June 2008 that the economy had grown during that quarter.[8] Official data also often cover only a small part of the actual economy. It is possible that, by giving policymakers access to real-time information covering a much larger portion of actual transactions, Big Data could improve the ability of fiscal and monetary officials to avoid policy errors and allow businesses to time their investments more accurately.[9]

Big Data is also having an enormous impact on international trade. Data flows are the fastest growing component of international trade.[10] Another McKinsey report found that global flows of trade, finance, people, and data increased world GDP between $250 billion and $450 billion each year.[11] This report also found that economies with more international connections received up to 40% more benefit than less connected economies.[12]

Private Sector Value and Potential

Big Data will have a disproportionate impact on many industries. A 2013 report commissioned by the Direct Marketing Association measured the size of the data-driven marketing economy (DDME), defined as the set of firms that produce marketing services focused on individual-level consumer data for marketing firms.[13] It found that in 2012, producers spent about $156 billion on these services, creating employment for about 676,000 people. The implication is that the buyers of this information derived at least this much value from it. Importantly, it found that roughly 70% of this value and employment depended on moving data between firms.

The study also found that the main benefit of the DDME was that it made marketing more efficient, allowing companies to avoid sending solicitations to individuals who are unlikely to buy their products and instead target prospective customers with offers that better match their needs and interests. A second benefit is that sellers are able to improve their effectiveness by matching specific marketing efforts with results. The DDME also reduces the barriers to entry for small manufacturers because it lowers the cost of obtaining and using high-quality consumer data. This benefit would not be available unless a robust market was allowed to exist in consumer data.

Because of their increased importance as an economic resource, restricting data flows can seriously hurt national welfare. A study by the European Centre for International Political Economy and the U.S. Chamber of Commerce concludes that implementation of the European Union's proposed General Data Privacy Regulation would reduce EU exports to the United states by between 0.6% and 1%, undoing much of the potential impact from the proposed Transatlantic Trade and Investment Partnership.[14] The negative results were reduced because the regulation would replace national data restrictions that are already in effect and allow for workarounds, such as model contract clauses and binding corporate rules to substitute for direct regulation. Eliminating these workarounds would have an even larger effect, reducing EU exports to the United States by 4.6% to 6.7% and EU GDP between 0.8% and 1.3%.[15]

Further emphasizing the importance of data mobility, the Omidyar Network recently released an economic analysis of adopting the type of Open Data policies discussed in Chapter 6 of this report. The study concludes that implementation of Open Data policies could boost annual income within the G20 by $700 billion to $950 billion.[16] Significantly, the benefits come in a wide variety of forms, including reducing corruption, improved workplace conditions, better energy efficiency, and a reduction in the regulatory costs associated with international trade.[17]

The quest to gather and use consumer data has also generated a large increase in Internet advertising. A recent McKinsey study found that these ads underwrote the delivery of a range of free Internet services that delivered significant benefits to Internet users.[18] The study estimated

that in 2010, these services generated a social surplus (the excess of benefits over costs) of €120 billion. Significantly, 80% of this surplus went to consumers. Consumers will only continue receiving these benefits so long as advertisers receive value from funding them.

Much of the data impact is and will continue to be in the information technology industry, as the demand for sensors, data storage, processing capacity, and software increases. McKinsey cites studies that global data generation will increase by 40% per year.[19] Nearly 80% of this is apparently copies of existing data.[20] From 1986 to 2007, data storage and computing capacity increased by 23% and 58%, respectively. Virtually all of this information is now in digital form, making it much easier to copy, analyze, transmit, and store.[21]

Then there is the impact on the labor market. Data analysis has been labeled "the sexiest job of the 21st century."[22] One estimate finds that there are already around 500,000 Big Data jobs in the United States.[23] Still, the McKinsey studies point to a serious shortage of managerial talent capable of understanding and acting on Big Data. Most visible are the data experts with advanced degrees in statistics, computer engineering, and other applied fields. McKinsey finds a national shortage of between 140,000 and 190,000 people. But just as serious is the shortage of 1.5 million managers and analysts in traditional jobs who are capable of asking the right questions about the data and acting on the answers. Also important are the line employees who must properly implement data strategies. In fact, the inability to find and keep workers with even moderate math and statistics skills is already placing limits on business profitability.[24] The demands of the data-driven economy will only exacerbate the current shortage of well-educated workers.

The impact on labor markets will not be totally positive. Some have expressed concern that a data-driven economy will eliminate jobs through a combination of automation and increased competition.[25] Technology has frequently produced highly disruptive changes to the economy, and the pace of these changes may well increase as a result of future advancements in information technology. Yet, an international study by McKinsey found that within small- and medium-sized enterprises, the Internet (and by implication, the data-driven economy) created 2.6 jobs for every 1 it eliminated.[26]

Finally, cloud computing has increased the power of the information system. The ability to lease cheap storage and processing power has two important economic impacts. First, it transforms a large fixed cost into a variable cost. Companies can avoid having to purchase and maintain their own data centers and write or purchase their own software programs and instead lease both on an easily scalable, as-needed basis. Sophisticated data strategies no longer require large up-front capital costs and deep expertise in computer maintenance. Second, even the smallest companies now have access to the fastest servers and most sophisticated processing power at affordable rates. By making it easier for all companies to enter new and existing markets, cloud computing should increase both the diversity and competitiveness of markets.

The Role of Data in the Economy

The challenge for businesses will be to find the necessary talent that allows them to discover the true causal relationships within the data. They must then use these relationships to implement profitable business strategies in ways that do not violate public expectations about the proper uses of data. Developing and implementing successful innovation based on data insights will often be the hardest challenge. And they must often do this in real-time, responding even as the causal relationships change.

Much of the value in Big Data is likely to come from combining Big Data with the Internet of Things. Cheap sensors and transmission capacity can be used to generate enormous amounts of fresh data, which can then be fed into a system capable of analyzing and acting on it to solve existing problems. Figuring out how to act on the resulting flow of information, however, may not be easy or cheap. Consider parallels to the introduction of electricity into manufacturing plants, which forced a significant reengineering of manufacturing activity as plants that were built to harness other forms of power tried to optimize the value of this new resource.

Companies can create value by using data to solve problems. Some problems are unsolvable in the sense that the data needed to solve them does not exist, anywhere. A good example at present is

how to cure late-stage cancers. For most, however, the challenge is that the data needed to solve a problem have often been difficult to collect or are mixed in with other, unrelated data.

The promise of the information system is that it makes it possible (and increasingly affordable) to collect the right information, process it into actionable knowledge, transmit that knowledge to the right person, and act on it. In doing so, it allows us to solve an increasing number of problems, including many that we had never thought of. Sometimes the border between solvable and unsolvable problems is fuzzy. For example, we currently do not often know much about why some students learn at different speeds. Yet, it may be that if we had more data about all students, including individual students' strengths and weaknesses, we could design more effective educational software.

Companies that want to use Big Data to improve their internal operations will need to identify the key problems currently holding them back. These might include poor inventory management, difficulty retaining customers and workers, or poor decision making. UPS has increased its profit margins by collecting detailed information about the location and performance of their large vehicle fleet.[27] Data can also help businesses spot previously unrecognized problems. Gathering data on the water use of individual households, for example, can help municipalities identify anomalies that deserve a closer look.

Companies seeking to develop new products or services need to identify unsolved problems among their customers. For instance, every car driver has occasionally become lost. Every parent has wondered whether their child is learning what she should. Every diabetic needs to know his blood sugar level. To be successful, companies will have to identify these challenges and needs and then figure out what data is needed to solve them, transmit it to the right place at the right time, and act on it, all in a way that is intuitive for the customer. One key rule of technology in general, and data in particular, is that it will sit idle if it is too difficult to use.

Overall, the greatest value will go to companies that can identify unmet problems or needs, both within their own operations and among their customers. It sounds easier than it is. Executives have difficulty imagining new management opportunities (such as Total Quality Management and Six Sigma),[28] and customers have difficulty articulating a need for products like the iPhone and FitBit. The prior infeasibility of collecting and using the necessary data is likely to have prevented the recognition of many problems. That is what made both Pandora and the Nest thermostat such innovative products.

Management theorists, including W. Edwards Deming, have broken down the process of continuous improvement into a four-part cycle: plan, execute, measure, and adapt.[29] Data, and the ability to understand it, is critical to this process. It allows people to compare planned results with actual outcomes and then adjust their future actions to reduce the gap between the two. The key challenge often lies in deciding what to measure and integrating the right information into a process designed to improve performance. Toyota and other companies, for example, used this to great effect in the 1980s. A key part of Toyota's process was collecting detailed information about the manufacturing process, including production rates and quality measures. The company then used this data to spot problems, identify their root causes, and implement lasting solutions. The result was a dramatic improvement in profit and market share because of better quality, lower costs, and shorter production cycles.[30]

Big Data transforms this process in several important ways. First, the time lags involved in collecting and analyzing data often imposed a significant delay between execution and measurement in the improvement cycle. For example, the negative effects of changes in maintenance procedures might not become apparent until machines begin to fail faster than normal. With cheap sensors and rapid transmission, companies increasingly have instantaneous insight not only into their own performance but also into the performance of their products long after they have left the factory. This allows for a closer connection between plan implementation and data response and permits companies to develop and produce further iterations much faster.

Second, increasingly granular data allows companies to improve performance through A/B testing. For example, by varying product layout slightly between two stores and then measuring daily traffic and sales in each location,

management can see which of the two variations leads to better results. Results show that minor changes in the layout of a Web landing page can increase customer inquiries by more than 300%. Which version is better, however, is not always obvious until one has the data.[31]

Third, the ability to store and process large amounts of data allows companies to search for subtle relationships between them. Whether it is Target analyzing the collective buying decisions of millions of shoppers, geneticists looking at the combination of millions of genomes and personal histories, or educators studying data on learning outcomes, in many cases, meaningful relationships between different factors do not become apparent until researchers have lots of data points to study. This is especially true with many factors, each of which slightly influences the probability of a given outcome. For instance, in Boston, Bridj looks at a large collection of data feeds from Google Earth, Foursquare, Twitter, Facebook, the Census, and other sources to figure out where commuters are and where they want to go. It then arranges temporary bus service to meet the demand. Rather than view these new routes as fixed, Bridj tries to respond to market changes.[32]

Yet, the mere presence of Big Data does not guarantee economic profits. In fact, firms may become misled because the link between more data and better outcomes is not perfect. As Tim Harford has pointed out,[33] enthusiasts for Big Data have made four important claims: (1) data analysis produces extremely accurate results; (2) every data point can be captured; (3) we do not need to understand why data are correlated; and (4) scientific models are unnecessary. At best, these are generalizations that, if taken for granted, can lead to poor business decisions.

A fundamental assumption is that data samples are unbiased. Although tests exist to detect and treat bias, they are not perfect, and companies may still find themselves using data that are not representative of the population they supposedly represent. It is often the case that the data produced by the exponentially growing number of 21st-century technologies do not accurately reflect the characteristics of the entire population. Users of Foursquare and Twitter constitute a discrete subset of the American population, as do the consumers who take the time to register their products online. The most engaged and demanding customers are likely to have tastes that differ from those of the average shopper. In other words, the data that is easiest to collect may not be the most valuable. Companies that base their marketing strategies on data generated by a subset of the market may find that their product appeals to only that minority.

Second, firms will continue to face a trade-off between carefully targeting high-value customers and focusing on too narrow a portion of the potential market. Big Data might help them distinguish between the two but only to a point. It will still be the case that efforts to selectively target only the most promising customers will miss a lot of potential business, while campaigns that more broadly target all potential customers will spend money on people who were never potential buyers.

The size and science of Big Data can easily lend a false sense of precision to the equations that come out of it. As an example, much has been made of the fact that Target once sent ads for baby products to a teenager before her father knew she was pregnant. To know exactly how significant this was, however, we would need to know what proportion of women who received those ads were actually pregnant. If Target sent them to everyone, the significance of the event would disappear.

The bigger data is, the more likely it is to contain spurious relationships. This is especially true when the data pile contains a large number of variables and managers are mining the data for correlations. The problem is that chance correlations often exist. Because these relationships are spurious, they are unlikely to persist going forward. Thus, basing a decision on these correlations is dangerous. The next data pile may also contain lots of correlated variables, but they are likely to be different from the first set. In the meantime, companies that relied on the first set of correlations will have aimed their resources in the wrong direction.

What is more, even if correlations between data are causal, there is no assurance that these relationships will continue into the future. Indeed, the key to business success may often lie in disrupting traditional business relationships and transcending product boundaries by offering innovative products that redefine the market and play to the strengths of a particular firm. Because

these products are new, there may be little data to guide executives.

For data to add value, its use must be properly inserted into an institutional setting. For example, Street Bump is an app that mobile users can download to help the city of Boston locate potholes. The app sends a signal to the city every time its owner passes over a pothole. The idea is ingenious, but it is only likely to make a real difference if identifying potholes is the binding constraint in the system. If without Street Bump potholes would be identified within two days anyway, then the added value may be small. If the potholes are otherwise identified by city workers who are sent around to look for them, the app may save money but only if the workers are let go or assigned to more productive work. Finally, if the real constraint on street repair is a shortage of funds, equipment, or labor, Street Bump may only add to the backlog of unfilled potholes and have little impact on actual street quality.

The Social Dimensions of Big Data

Any resource's value depends upon the social, legal, political, and economic environment surrounding it. For instance, the progress of fracking in the United States compared to other countries is at least partially the result of the facts that: U.S. landowners own and can convey the mineral rights to their property; the energy industry had a large number of small, private companies that had to innovate or die once U.S. production started declining; and companies could easily raise money in open, sophisticated capital markets.[34] Some of the primary factors that affect the economic potential of Big Data include ownership rights, security concerns, and the level of transparency surrounding data's collection and use.

Ownership

Ownership of data raises important social issues. It will often be the case that the ownership of data will not be linked to its possession. For example, individual users of social networking sites are likely to believe that they own the content they generate, irrespective of what the Terms of Agreement state. Those views are likely to drive legislation if the companies holding the data fight this perception too much. This concept of ownership might also extend to the full extent of what McKinsey calls MyData,[35] which it defines as all of the data generated about a person, regardless of whether he is aware of it and whether he is the one collecting and storing the data.

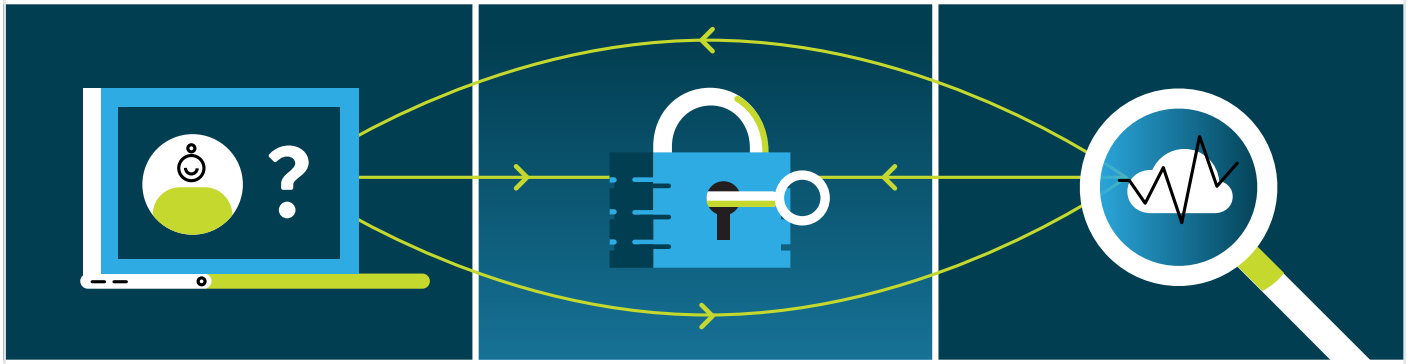
MyData can therefore consist of much more than a Facebook or Twitter account. It includes one's personal shopping history, be it on websites or in brick and mortar stores. Yet, it increasingly also covers a wide variety of other data, including: health data that users generate about their weight, sleeping patterns, and diets; volume and timing of utility consumption; and personal location. A growing number of providers are likely to compete in giving individuals greater access to this information and helping them understand and act on it.

Security

Security and privacy are intimately linked: the more private the data, the more securely people expect it to be held. We are coming to realize that security concerns represent a significant potential liability for firms. A large data breach can cost a company a great deal in terms of money, customer loyalty, and regulatory involvement. Companies that maintain possession of large amounts of consumer data, especially data that can be individually identified, need to implement the best security

"ANY RESOURCE'S VALUE DEPENDS UPON THE SOCIAL, LEGAL, POLITICAL, AND ECONOMIC ENVIRONMENT SURROUNDING IT."

THE SOCIAL DIMENSIONS OF BIG DATA



OWNERSHIP

Who "owns" social and financial digital data about an individual? In most countries, the answer is the company or organization that collected it. But policies such as the "Right to be Forgotten" in the European Union are challenging that concept by granting new data rights to individuals.

SECURITY

The public is willing to trust companies and organizations with financial, private, and often very personal data because they expect it to be held securely. Data breaches undermine that trust, invite further regulation, and effectively stifle innovation.

TRANSPARENCY

The highly personalized service experience enabled by Big Data will eventually lead to greater data transparency. The market will reward brands and stores that combine customized shopping with a reputation for fair, secure, and open data policies.

measures. This may require enough expertise to justify involving outside companies that specialize in data security and storage. To some extent, the problem is self-correcting; the boards of every major company are now focused on how to prevent a data breach.[36]

Transparency

The growing use of Big Data is also likely to result in greater transparency. Companies have already put a lot of effort into identifying and tracking their best and worst customers. Big Data could allow them to have a personal relationship with each customer, tailored to that customer's needs and individual preferences. Each individual may increasingly appear to both companies and governments as a unique entity with a full history of purchases, payments, income, goals, and more. This greater insight into people (and companies) can help providers deliver products that meet their customers' deeper needs by identifying and responding to individual characteristics. It should also make it easier to form longer-lasting relationships, where value is based less on price and more on personal meaning to the customer.

It is true that transparency is likely to pressure profit margins by making it easier for customers to comparison shop, but it will also make it easier for buyers to verify a seller's quality claims, strengthening the market for higher valued-added products. As the economy grows and higher incomes allow consumers to search for more personally rewarding experiences, the willingness to pay a premium for quality and tailored products and services should increase.

In the same vein, companies are becoming increasingly transparent to consumers. With the aid of social networks, third-party data aggregators, and mobile technology, customers have access to much more information about the quality, structure, and ethics of the companies with which they interact. They will increasingly be able to trace their food back to the farm or factory from which it originated. Producers may find themselves competing with each other to give consumers an open view of their processes, perhaps even by placing cameras in their production centers so that customers can verify claims of safety. This trend could make each company within a given

supply chain (particularly those whose brand is attached to the product) more responsible for the performance of the entire supply chain.

Conclusion

Financial derivatives are a standard business tool for managing a wide variety of risks. When they were first adopted, however, many companies were focused on their novelty rather than on the business case for using them. As a result, a number of firms suffered large losses. It took time and hard lessons for companies to learn how to properly integrate derivatives into normal business operations. The same learning process is already taking place with Big Data. Like financial innovation, it is an always-evolving concept, requiring constant education and adaptation. That said, it offers huge rewards for those who succeed in using it to create value for customers through better products and services.

By increasing the availability of data and reducing its cost, the data-driven economy promises a smarter, more efficient world. But it will not solve all problems. The path between gathering data and acting on knowledge involves many steps, not all of them subject to improved technology. Many of today's most pressing problems involve a difference of values rather than a disagreement about facts.

Big Data represents one of the largest untapped resources yet. Thanks to continued advancements in information technology, it is finally being tapped. Together with the rise of the Internet of Things, it constitutes a general purpose technology. Such technologies have broad impacts on the economy and society. The full impact from Big Data and related technologies will be spread out over several decades. This is partially because use of any new resource or technology often requires a significant transformation of the status quo. It takes time for people to think of new ways to use the resource and implement the necessary changes. Nevertheless, the promise of Big Data is transformative and its economic impact expansive, cascading, and world changing.

ENDNOTES

1 *See*, Fred Krupp, "Don't Just Drill, Baby—Drill Carefully: How to Make Fracking Safer for the Environment," *Foreign Affairs,* May/June 2014.

2 Michael Mandel, "The Data Economy is Much, Much Bigger Than You (and the Government) Think," *The Atlantic*, 25 July 2013.

3 *See,* Michael Chui, Markus Löffler, and Roger Roberts, "The Internet of Things," *McKinsey Quarterly,* March 2010.

4 Organization for Economic Cooperation and Development, "Exploring Data-Driven Innovation as a Source of Growth," *OECD Digital Economy Papers*, No. 222, June 2013.

5 The seven industries are education, transportation, consumer products, electricity, oil and gas, health care, and consumer finance. *See*, James Manyika et al., "Open Data: Unlocking Innovation and Performance with Liquid Information," *McKinsey Global Institute*, Oct. 2013.

6 Laurie Sullivan, "Wal-Mart RFID Trial Shows 16% Reduction In Product Stock-Outs," *InformationWeek,* 14 Oct. 2005.

7 Erik Brynjolfsson, Lorin M. Hitt, and Heekyung Hellen Kim, "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?" 22 April 2011.

8 "Gross Domestic Product: First Quarter 2008," *U.S. Department of Commerce Bureau of Economic Analysis*, 26 June 2008.

9 Michail Skaliotis and Ceri Thompson, "Big Data: An Opportunity to Boost Analytical Capacities," *SciDevNet*, 15 April 2014.

10 Michael Mandel, "The Data Economy is Much, Much Bigger Than You (and the Government) Think," *The Atlantic*, 25 July 2013.

11 James Manyika et al., "Global Flows in a Digital Age: How Trade, Finance, People, and Data Connect the World Economy," *McKinsey Global Institute,* April 2014

12 Ibid.

13 John Deighton and Peter A. Johnson, "The Value of Data: Consequences for Insight, Innovation and Efficiency in the U.S. Economy," *Data-Driven Marketing Institute*, 14 Oct. 2013.

ENDNOTES CONTINUED

14 European Centre for International Political Economy, "The Economic Importance of Getting Data Protection Right: Protecting Privacy, Transmitting Data, Moving Commerce," *U.S. Chamber of Commerce*, March 2013, 13.

15 Ibid., 15.

16 Lateral Economics, "Open for Business: How Open Data Can Help Achieve the G20 Growth Target," *Omidyar Network,* June 2014.

17 Ibid., xiv-xv.

18 McKinsey and Company, "Consumers Driving the Digital Uptake: The Economic Value of Online Advertising-Based Services for Consumers," *IAB Europe*, Sept. 2010.

19 James Manyika et al., "Big Data: The Next Frontier for Innovation, Competition, and Productivity" *McKinsey Global Institute*, May 2011, 16-17.

20 Ibid., 19.

21 Ibid., 16-17.

22 Thomas H. Davenport and D.J. Patil, "Data Scientist; The Sexiest Job of the 21st Century," *Harvard Business Review*, Oct. 2012.

23 Michael Mandel, "Where are the Big Data Jobs?" *Progressive Policy Institute,* May 2014.

24 Timothy Aeppel, "Bringing Jobs Back to U.S. is a Bruising Task," *Wall Street Journal,* 25 June 2012.

25 Erik Brynjolfsson and Andrew McAfee, *Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy,* (Digital Frontier Press, 2012).

26 Matthieu Pélissié du Rausas et al., "Internet Matters: The Net's Sweeping Impact on Growth, Jobs, and Prosperity," *McKinsey Global Institute*, May 2011, 21.

27 Mary Schlangenstein, "UPS Crunches Data to Make Routes More Efficient, Save Gas," *Bloomberg,* 13 Oct. 2013.

28 Total Quality Management and Six Sigma were disciplined methods of improving quality within a production process. Both methods stressed careful observation of existing processes, the collection of large amounts of performance data, and investigations to discover the root causes of production problems.

29 W. Edwards Deming, *The New Economics for Industry, Government, Education* (MIT Press, 1993).

30 James P. Womak, Daniel T. Jones, and Daniel Roos, *The Machine That Changed the World: The Story of Lean Production – Toyota's Secret Weapon in the Global Car Wars That is Now Revolutionizing World Industry* (Scribner, 1990).

31 See, James Gardner, "12 Surprising Test Results to Make You Stop Making Assumptions," *Unbounce*, 19 Sept. 2012.

32 Katharine Q. Seelye, "To Lure Bostonians, New 'Pop-Up' Bus Service Learns Riders' Rhythms," *The New York Times*, 4 June 2014.

33 Tim Harford, "Big Data: Are We Making a Big Mistake," *Financial Times*, 28 March 2014.

34 Robert A. Hefner, III, "The United States of Gas: Why the Shale Revolution Could Have Happened Only in America," *Foreign Affairs,* May/June 2014.

35 Manyika, "*Open Data,*" 4.

36 Danny Yardon, "Corporate Boards Race to Shore Up Cybersecurity," *Wall Street Journal*, 29 June 2014.

CHAPTER

2

ABOUT THE AUTHOR

Leslie Bradshaw is a managing partner at Made by Many, a product innovation company with offices in New York City and London. Named one of the "Most Creative People in Business" by *Fast Company* in 2013, Bradshaw's areas of focus are interpreting and visualizing data, knowing what it takes to build and grow companies, and how to create lasting business impact through innovation. A graduate of the University of Chicago and contributor to *Forbes,* Bradshaw led her first company to the Inc. 500 list twice for revenue growth experienced during her tenure.

THE GREAT DATA REVOLUTION

BY LESLIE BRADSHAW

Key Takeaways

The human brain is still the most valuable data analytics tool. Even the most advanced forms of number crunching and correlation recognition are useless without contextual application and analysis.

Data literacy is more important than ever for policy makers, business leaders, and entrepreneurs and citizens. Core mathematics curriculum in primary schools should emphasize not only the building blocks of algebra and calculus but also critical reasoning and data visualization.

It is ultimately more productive for policy makers and business leaders to think about Big Data in terms of its functionality—facilitating data-driven innovation—rather than its dictionary definition.

Media coverage of the Big Data revolution tends to focus on new technology developments in data storage and new business opportunities for social analytics and performance management. Alongside these tech-sector updates, a parallel media narrative has focused on the public's concern over data collection practices.

Market success stories and accentuated privacy concerns are both important and deserve the full attention of business leaders and policymakers. Yet, we should recognize that much of what we read in the press about data is only the protruding tip of the proverbial iceberg. The layer of information generated by Big Data permeates through not only the densely linked networks of business, finance, and government, but into all aspects of quantitative research and scientific inquiry. Big Data is thus much more than an impressive technological development—it is a new framework for understanding and interacting with the world around us.

Big Data offers a new era of learning, where we can investigate and analyze a larger body (or the entire body) of information about a subject and gain insights that were inscrutable in smaller samples. Big Data is *already* reframing critical questions about the processes of research, best practices for engagement with all categories of digital data, and the constitution of knowledge itself.[1] Moreover, while business and consumer stories occupy the headlines, some of the most promising applications of the technology are found in nonprofit work, good governance initiatives, and especially, scientific research.

This chapter colors outside of the lines of familiar Big Data narratives and addresses some of the underreported and less understood aspects of the phenomenon. It explores the value-added aspects of Big Data that make it more than its component parts, and it makes the case that data is best conceptualized—and applied—as a complementary extension of human ingenuity. Computers can crunch numbers, but when it comes to contextualizing and applying that analysis, only a human mind will suffice.

This chapter also emphasizes the importance of data literacy as both an organizational best practice and a core curriculum. It explores how Big Data is being used by nonprofits and universities to alleviate some of the world's most pressing problems (such as disease control,

environmental issues, and famine reduction) and considers how Big Data can be applied at the micro-level for individual optimization. The chapter concludes with options for reframing the Big Data debate into a benefits-oriented discussion of data-driven innovation.

A Quantitative Shift

One challenge we find when talking about Big Data is that the term is often described by way of anecdotal example (rather than formal denotation). Big Data might be "Google's satellite mapping imagery," "the streaming financial data supporting Wall Street," or even "all of the people on Facebook," depending on who you ask. Yet, do all these examples really represent the same thing? How big does data have to be before it can be considered Big Data? Is Big Data really a "thing" at all—or is it also a process? Can we effectively promote the benefits of Big Data when we can't even agree on what Big Data is?

Most observers would agree that Big Data is a broad, catch-all term that captures not only the size of particular datasets but also advances in data storage, analytics, and the process of digitally quantifying the world. Big Data may be a nebulous term, but that doesn't mean it is useless. We commonly speak about equally vague technological terms like "social media," "cloud computing," and even "the Internet" without prefacing every remark with peer-reviewed and linguist-approved qualifications. Big Data is more of a *dynamic* than a thing, but the different facets and technology developments that reflect that dynamic are well known—leading to a multitude of anecdotal examples.

What makes Big Data so useful? It's a complicated—and highly contextual—question, but a simple response really does begin with the defining descriptive attribute: volume. For analytical purposes, more data tends to produce better results. Peter Norvig, an artificial intelligence expert at Google, provides an illustrative analogy in his presentation on "The Unreasonable Effectiveness of Data."[2]

Norvig notes that a 17,000-year-old cave painting effectively tells its audience as much about a horse—a four-legged, hoofed mammal with a thick mane—as any photograph. While drawing the animal with dirt and charcoal is a much slower process than snapping a picture, the information conveyed is fundamentally the same. No matter how advanced the technology that produces it, a single piece of data will always contain a limited amount of both implicit and contextual information. Capturing consecutive images of a horse in the form of a video, however, produces a much fuller assessment of how the animal moves, behaves, and interacts with its environment. Even a modest quantitative shift in the data allows for a far more qualitatively rich assessment.

In the Big Data era, we can not only capture a series of videos of a horse; we could capture the animal's *every* movement for hours, days, or weeks. Before Big Data processing programs, organizations could not effectively analyze all of the data points they possessed or collected about a particular phenomenon. That was why accurate, representative sampling was so important. Today, it's not only possible, but preferable to pull and analyze *all* of the data.

Volume, however, isn't the whole story. Although many of the datasets identified in press accounts are staggeringly large (such as the 200-terabyte dataset for the 1000 Genomes Project, cataloging human genetic variation), other datasets lumped in with this trend are not nearly as extensive. Big Data is ultimately less about the size of any one dataset than: (1) a capacity to search, aggregate, and cross-reference an ever-expanding ecosystem of datasets, which include the incomparably large and the proportionally small; and (2) an ability to render previously qualitative research areas into quantitative data.

An excellent example of the latter is Google's Ngram Viewer. In 2004, Google began scanning the full text of the world's entire body of books and magazines as part of its Google Print Library Project. This digitization effort was eventually folded under the Google Books label, which today encompasses more than 20 million scanned books. The Ngram Viewer allows users to search through 7.5 million of these books (about one-seventh of all books ever published) and graph the frequency with which particular words or phrases have been used over time in English, Chinese, Russian, French, German, Italian, Hebrew, and Spanish-language literature.[3]

Time referred to the project as perhaps the "closest thing we have to a record of what the world has cared about over the past few centuries."[4]

Ngram Viewer: Bridging Textual and Quantitative Analysis

Google's Ngram Viewer allows users to search the full text of 7.5 million digitized books published over the past 200 years. The program, based on a prototype called "Bookworm," was created by Harvard doctoral candidates Jean-Baptiste Michel and Erez Aiden and MIT programmer Yuan Shen. With help from Shen, Michel and Aiden set out to create a "microscope to measure human culture" and to "identify and track all those tiny effects that we would never notice otherwise." Their book, *Uncharted: Big Data as a Lens on Human Culture*, spotlights some of the fascinating results from the authors' analysis of centuries of word usage, including attempts to quantify the impact of censorship.

The term "Tiananmen," for instance, soars in English-language publications after the Tiananmen Square Massacre. In Chinese literature, however, the term receives only a brief blip of interest. How can this be? According to the authors, after the massacre, Chinese officials carried out a remarkably effective campaign of censorship and information suppression to scrub out negative references to the incident. Although the massacre is one of the central events in modern Chinese history, nobody in China (outside of a select few government officials) is allowed to discuss it, at least not in print. The incident simply doesn't exist in the historical record—a testament to what Michel and Aiden call the "brutal efficiency of censorship in contemporary China."

The authors also use Ngram to offer quantitative insights into the half-lives of irregular verbs, the origin of the word "chortle" (Lewis Carroll's nonsense poem "Jabberwocky"), and the criteria for lasting fame—or at least notoriety, as Adolph Hitler remains the most referenced figure born in the past two centuries, and Joseph Stalin and Benito Mussolini are not far behind.

See:
Erez Aiden and Jean-Baptiste Michel, *Uncharted: Big Data as a Lens on Human Culture* (New York: Riverhead, 2013).

A more modern example is the proliferation of data-driven journalism outlets. While "data" journalism in the broadest sense has been around for ages (think political polls and census analysis), the influx of user-friendly statistical software, easily accessible spreadsheets, and data-curious reporters has transformed a niche concentration within newsrooms into a whole new type of journalism.[5]

The most prominent example is FiveThirtyEight. Despite some early stumbles, the site has found a loyal audience through compelling political analysis (especially election predictions and polling criticisms), statistics-heavy sports features (site founder Nate Silver got his start in baseball sabermetrics), and even irreverent lifestyle features (anecdotal reporting). This quantitative approach is reflective of a larger shift in how we read, learn, and process information. Today's readers are no longer satisfied with two-dimensional news—they want strong reporting and analysis presented in an engaging, easy-to-digest (read: mobile-optimized) presentation. The sort of interactive infographics, explanatory videos, and multi-platform features that are thriving online simply weren't possible in the old days of print. News organizations are responding to demand for these stories by actively recruiting reporters with a statistical background and designers skilled in data visualization.

From Raw Data To Useful Information

It is easy to imagine Big Data as a massive Excel spreadsheet just waiting for somebody to hit "sort," but that's not quite right. In many cases, Big Data is closer to the unsorted mess of memories and factoids floating around in our heads. This wide array of information and details can only be processed through the metadata (see Chapter 4) that facilitates dense linkages and logical pattern recognition. Changing raw data to actionable information, then, requires a full understanding of context.

The functional relationship between data and information is detailed in the Data-Information-Knowledge-Wisdom (DIKW) pyramid, a heuristic device brought to prominence in the late 1980s by organizational theorist Russell Ackoff.[6] The pyramid explains that information is typically defined in terms of data, knowledge in terms of information, and wisdom in terms of knowledge. Theorists contest some of the finer points of this logical progression—especially the distinction (if there is one) between wisdom and knowledge—but as a general framework, the DIKW pyramid remains useful in demarcating analytically fuzzy concepts.

When critics challenge some of the claims surrounding Big Data, they are usually targeting misunderstandings about the relationship between information and data. For example, *Wired* editor

Chris Anderson famously claimed that with enough data, the numbers would "speak for themselves" and "make the scientific method obsolete."[7] Anderson's wide-eyed assessment (which was widely mocked, even by Big Data practitioners) was incorrect because he failed to recognize that data is useless without context, theory, and interpretation.

In a great *New York Times* op-ed, New York University's Ernest Davis and Gary Marcus point out that a Big Data analysis of the crime rate in all 3,143 counties in the United States between 2006 and 2011 might reveal that the declining murder rate is strongly correlated with the diminishing market share of Internet Explorer.[8] A similarly comprehensive analysis of autism diagnosis cases and organic food consumption might reveal a statistically significant correlation. Big Data can produce endless examples of such correlations but is thus far ineffective at determining which correlations are meaningful.

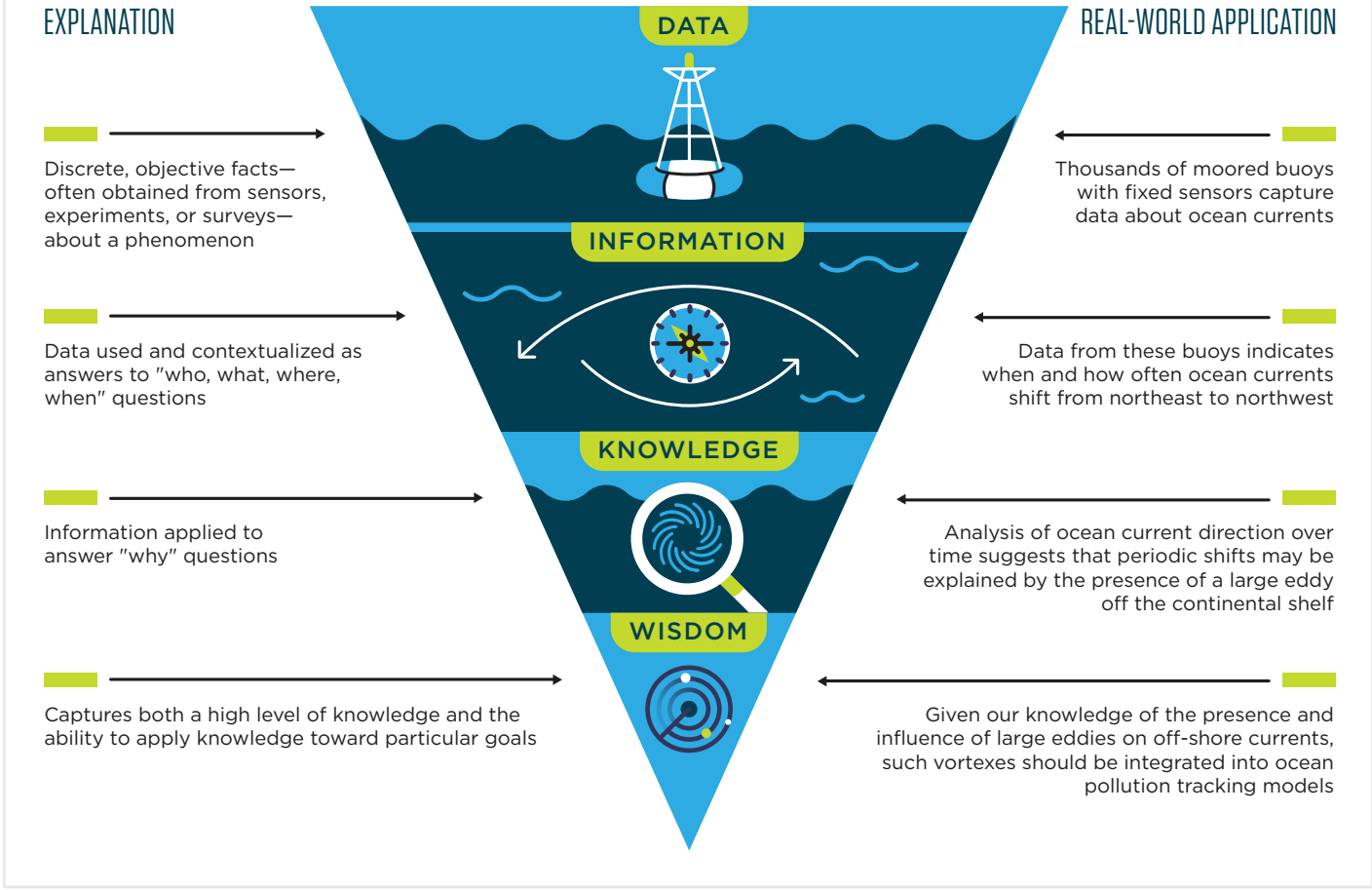
Simply put, even the most advanced forms of number crunching and correlation recognition are useless without contextual application and analysis. In this area at least, even the fastest computers and most powerful analytic applications still trail the human mind, which is uniquely capable of making such connections. Writing in *VentureBeat*, ReD's Christian Madsbjerg and Mikkel Krenchel note that while computers excel at following narrowly defined rules, only the human brain is capable of reinterpreting, reframing, and redefining data to place it within a big picture.[9]

Until computers are able to "think" creatively and contextually—or at least are able to mimic such cognitive functioning—the human brain will remain a necessary conduit between data analysis and data application, which is reassuring. Not only will Big Data *not* make humanity obsolete, advanced technology will make our most creative faculties more relevant than ever. Indeed, Big Data is the latest era-advancing piece of technology (not unlike world-changing innovations such as the printing press, steam engine, and semiconductor) that can be used to expand ontological horizons and scientific capabilities.

That said, a large caveat is in order: countless business headlines and anecdotal examples suggest that humans are just as capable of drawing the wrong conclusions from data as the correct ones. This is why data literacy is so important—both as an organizational best practice and as an educational praxis. Without the ability to understand and communicate data correctly, we may end up collecting the wrong data, ignoring the right data, failing to apply the data (or applying it incorrectly), extracting the wrong meaning from it, or twisting the results to support our preconceptions.

"UNTIL COMPUTERS ARE ABLE TO "THINK" CREATIVELY AND CONTEXTUALLY—OR AT LEAST ARE ABLE TO MIMIC SUCH COGNITIVE FUNCTIONING—THE HUMAN BRAIN WILL REMAIN A NECESSARY CONDUIT BETWEEN DATA ANALYSIS AND DATA APPLICATION."

BIG DATA'S FUNCTIONAL VALUE EXPLAINED: THE DIKW PYRAMID



Cultivating a Data-Literate Workforce

Most contemporary usages of the term "literacy" refer not only to the ability to read and write but also to the skills required to think critically about *how* something is written and *what* it may represent. More sophisticated definitions also capture the ability to apply these skills for personal development and social transactions.[10] For example, policymakers recognize that an elementary level of "computer literacy" has essentially become a prerequisite for participation in modern society. As such, computer education has been seamlessly integrated into grade school curricula, and government programs and civil society initiatives have emerged to bring older adults up to speed.

The term "data literacy" captures a number of core deductive logic and statistical analysis skills that predate the shift to digital, but in the Big Data era, these abilities are more critical than ever. Data literacy is defined primarily by its active functional component—the ability to convert data into valuable and usable information. Retailers, marketers, and tech leaders have been ahead of the curve on this, transforming themselves into data-driven innovators through sizable investments in new technology and training.[11] Universities have followed suit. Business analytics is gaining popularity as a curriculum focus within prominent MBA programs, while schools like Columbia University, Northwestern, New York University, and Stanford have launched quantitative studies and data mining programs.[12] These courses of study prepare students to:

- Use statistical methods to extract patterns and trends from large datasets;
- Develop and use predictive models and analytics;
- Understand and use strategic decision-making applications; and
- Communicate findings in practical business language.

The increasing number of data-literate college graduates and business professionals is a good sign—although McKinsey still expects the United States to have a shortage of up to 190,000 data scientists by 2020.[13]

Yet, colleges and business are not the only sources of a data-literate workforce. Learning how to crunch numbers and use the results to tell stories with words and visuals can (and should) start as early as elementary school. An over-reliance on calculators, computers, and text, however, threatens some of our most innate and powerful tools that lead to data literacy. Thus, teaching critical reasoning and visual storytelling skills is critical throughout K-12 and college education, as well as in the professional world.

In a broader societal sense, data literacy should reflect a more *passive* level of competency and awareness among all people, much the way most people have a working knowledge of personal credit ratings or online banking. The proliferation of digital data impacts all of us, and it shouldn't require a master's degree in computer science for citizens and consumers alike to understand what sort of data is being collected and how and why this data is being analyzed and applied.

The private sector has an important role to play here as well. One of the reasons privacy concerns are raised in the Big Data discussion is that consumer data-collection practices remain opaque and poorly understood, even by practitioners. Forward-thinking businesses that "get" data would be well-advised to translate organizational data literacy into public-facing data resources. By proactively taking on consumer education, companies are able to responsibly pursue policies that benefit and also protect consumers. Businesses that don't understand—or deceitfully mask—their own data usage policies might best pull back and reevaluate. Making policies known, clear, and uncomplicated is a best practice in a data-driven, increasingly data-literate world.

Data for Development

News coverage of Big Data is most prominent in the business section of the Sunday paper, where readers find numerous stories detailing the newest tech developments from IT leaders, online giants, and big-box retailers. These pieces are always worth a read, but the science, health, and weather sections hold articles that reveal much broader, more altruistic uses of Big Data. For example:

The Ocean Observatories Initiative recently began constructing a Big Data-scale cloud infrastructure that will store oceanographic data collected over the next 25 years by distributed sensors. The program will provide researchers with an unprecedented ability to study the Earth's oceans and climate.[14]

Flatiron Health, a Big Data startup that consolidates cancer treatment records to offer practitioners a clearer and centralized overview of patient needs, recently raised $130 million in funding from some big name backers. The company plans to create the world's largest pool of structured real-world oncology data.[15]

Monsanto recently acquired The Climate Corporation, a San Francisco-based company that maintains a cloud-based farming information system that gathers weather measurements from 2.5 million locations every day. Climate Corporation uses this trove of weather data to help farmers cope with weather fluctuations.[16]

All of these examples underscore the idea that Big Data isn't limited to big business. Indeed, data-driven innovation has already been institutionalized within Harvard's Engineering Social Systems (ESS) program, where researchers are looking to census data, mobile phone records, and other newly available digital datasets to provide insights about the causal structure of food shortage in Uganda, the necessity of transportation planning in Rwanda, and the complex behavior of human societies everywhere.

The ESS program is part of a growing consortium of nonprofits, government agencies, universities, and private companies that have been given the label "Big Data for development." Datakind is another bright star in that constellation. The New York-based nonprofit was created by *New York Times* R&D labs team member Jake Porway as a way to bring together data scientists and technology developers with civil society groups in a pro bono capacity. Porway recognized that while many nonprofits and social ventures accumulate large datasets about issues relevant to their missions, they often lack the technology resources and skills to perform analytics.[17] Datakind started

with local hackathon events but was soon working with the World Bank, Grameen Foundation, and the Red Cross to address problems ranging from fire prevention to good governance. The group now organizes data-dive events across the globe and will soon offer fellowships for longer term engagements.

Another great example is Global Pulse, a UN initiative that develops critical connections between data mining and humanitarianism. The organization uses real-time monitoring and predictive analytics to locate early warning signs of distress in developing countries. Global Pulse scans cell phone activity, social networking sites, and online commerce platforms for signals of near-future unemployment, disease, and price hikes, thus allowing for more rapid responses from humanitarian groups. The personal nature of this data does, of course, bring up privacy concerns, but Global Pulse's analysis does not identify specific individuals or even groups of individuals. Rather, the organization looks at large datasets of anonymized, aggregated data—much of it Open Data, discussed further in Chapter 6—that can provide a sense of "how whole populations or communities are coping with shocks that can result in widespread behavioral changes."[18]

Big Data is not only driving how nonprofits operate; it is also dictating how they receive funding. The increasing amounts of public domain and voluntarily provided information about charities, nonprofits, and related-tech ventures can help donors and investors channel dollars to the organizations that are most effective at fulfilling their objectives. Such assessments can be further facilitated by applications that collate and update data from ongoing evaluations, common performance measures, and qualitative feedback. A recent *Wall Street Journal* article speculates about an ROI-optimized world where "foundations will be able to develop, assess and revise their giving strategies by pulling information from community surveys, organizational reports, and an up-to-date 'ticker' of other philanthropic giving."[19]

This "Future of Philanthropy" is already happening. The Knight Foundation—which has emerged as the go-to philanthropic organization for funding "transformational ideas"—recently partnered with data analytics firm Quid to produce a detailed analysis of the financial investments that support "civic tech"-related ventures.[20] "Civic tech" is something of a catch-all category that captures startups, nonprofits, and new technologies that focus on improving the health and vitality of cities. This ecosystem of established operations and new ventures is so large that it was previously difficult (if not impossible) to determine, in a schematic sense, precisely from where funding was coming and the results it was producing.

Quid's approach allowed the Knight Foundation to map out the field through semantic analysis of private and philanthropic investment data. This analysis revealed that the civic tech field has exploded over the past decade, growing at an annual rate of 23% from 2008 to 2012. Quid identified 209 unique civic tech projects within that landscape. Peer-to-peer projects—such as Lyft (an app that facilitates ridesharing) and Acts of Sharing (which addresses all aspects of collaborative consumption)—attracted the vast majority of investment, followed by clusters of ventures related to neighborhood forums, community organizing, and information crowdsourcing. The aim of the analysis was not simply to sketch out the existing civic tech investment ecosystem but to help guide its future development.

Data for Good

Big Data encompasses not just the hardware and software advancements needed to work with data on a large scale but also the *process* of quantifying the world around us into observable and manageable digital data. Mayer-Schönberger and Cukier refer to this transformation as "datafication," and it is occurring constantly throughout the technology sector and at all levels of government and business.[21] It is well-established that this process also extends into our personal lives. It is nearly impossible to proceed through a normal day without leaving behind a digital trail of online activities (e.g., Amazon purchases; Netflix viewing). Some marketing firms and tech companies are even deploying anthropologists into natural social settings to further quantify (via sophisticated preference rankings) those few interactions that are not mediated through technology, such as our communal exchanges with one another and our impulsive interactions with branded products and new gadgets.[22]

There are generally two responses to our increasingly quantified lives. The first response is to push back. European Courts, for example, continue to recognize users' "Right to be Forgotten"—effectively placing the onus on the online giants (e.g., Facebook) to remove damaging personal information from search results when requested by wronged parties. Even in a world without social media and negative Yelp reviews, however, individuals would still generate an enormous amount of digital data by using credit cards, phone apps, and keycards. Meanwhile, marketers would still send out "individualized" coupons and e-mails based on circulated consumer profiles and publicly available data. In other words, no amount of pushback will stop the data-generating activities individuals perform every day, nor will it degrade the business advantage in analyzing available data and applying insights gleaned from it.

The second response to a quantified existence is that if it is going to happen, we might as well harness it in positive ways for our personal use, such as aiding in things like time management, career choices, weight management, and general decision making. For example, it is easy to begin keeping a detailed log of hours spent working, hours spent traveling, hours spent relaxing, and even miles logged on the treadmill. Numerous gadgets and software programs facilitate this personal quantification. The Up fitness band from Jawbone, for example, is designed to be worn 24 hours a day, 7 days a week. When used with the accompanying application, the device can collect data on calories consumed, activity levels, and rest patterns. Up allows users to analyze daily activity to see when (and for how long) they were most active or most idle. Maintaining quantitative data about our professional and personal routines can help us achieve a qualitatively better work-life balance. This example is indicative of larger Big Data trends that are breaking down the quantitative-qualitative barrier and transforming the way we interact with the world around us.

Unfortunately, the potential in Big Data is endangered by current frameworks that have a tendency to either over-complicate the topic and make it inaccessible to non-scientific audiences or create uneasiness around the topic by emphasizing privacy concerns. As George Orwell argues in his famous 1946 essay "Politics and the English Language," "An effect can become a cause, reinforcing the original cause and producing the same effect in an intensified form, and so on indefinitely."[23]

In other words, for businesses and policymakers, the way we talk about Big Data will define its use and either unleash or limit its value. To get away from this, it is much more productive to think about Big Data in terms of what it can and is enabling in every industry: innovation. To this point, it is widely recognized by policymakers and the business community that many of the most critical sectors of the economy are reaping the benefits of data-driven innovation. The healthcare industry uses digitized patient records to create more cohesive patient care between facilities; financial services use Big Data-enabled monitoring software for more accurate (and real-time) market forecasting; and public administrators use Open Data to increase transparency and facilitate more effective feedback loops. Drawing attention to these examples of how data drives innovation (and by consequence, economic growth) is far more beneficial than focusing on the size of the data, the processing power required to analyze it, and particularly, the rarely seen (though often hyped) negative ramifications for consumers.

The way we talk about Big Data can educate and clarify the dynamic through a results-oriented policy lens. Helping policymakers view Big Data from this big-picture perspective is important, and it better contextualizes benefits for individuals, organizations, and economies. Undue regulation may inadvertently hamper the technology's development and diminish near-future benefits. As argued in the Global Information Technology Report in 2014: "Decisions that affect data-driven innovation are usually focused on the problems of privacy and data protection, but fail to consider economic and social benefits that regulation could preclude."[24]

As with any transformational moment in business, there will be leaders and followers. Integrating Big Data thinking across the public and private sectors will not only benefit the bottom line for the companies who figure it out, but it will also benefit consumers, as they will be more informed and thus better able to navigate the Big Data landscape and enjoy all the benefits it offers. The companies that lead the way will therefore have a competitive advantage for reasons that span from creating greater internal efficiencies around usage to external impacts experienced by having more insights into and abilities to serve their customers.

ENDNOTES

1 Danah Boyd and Kate Crawford, "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," *Information, Communication & Society,* 15 no. 5 (2012): 662-679.

2 Alon Halevy, Peter Norvig, and Fernando Pereira, "The Unreasonable Effectiveness of Data," *IEEE Intelligent Systems Magazine,* 24, no. 2 (2012): 8-12.

3 Jon Orwant, "Ngram Viewer 2.0," *Research Blog*, 18 Oct. 2012.

4 Harry McCracken, "The Rise and Fall of Practically Everything, as Told by the Google Books Ngram Viewer," *Time*, 16 Jan. 2014.

5 Roger Yu, "Booming Market for Data-Driven Journalism," *USA Today*, 17 March 2014.

6 Russell Ackoff, "From Data to Wisdom," *Journal of Applied Systems Analysis,* 16 (1989).

7 Chris Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," *Wired,* 23 June 2008.

8 Gary Marcus and Ernest Davis, "Eight (No, Nine!) Problems With Big Data," *The New York Times*, 6 April 2014.

9 Mikkel Krenchel and Christian Madsbjerg, "No, Big Data Will Not Mirror the Human Brain—No Matter How Advanced Our Tech Gets," *VentureBeat,* 17 Nov. 2013.

10 Li Wei, ed., *Applied Linguistics* (Hoboken, NJ: Wiley-Blackwell, 2013).

11 In the mid-1990s, for example, British supermarket chain Tesco partnered with computer science firm Dunnhumby to establish Tesco's Clubcard, which allowed Tesco to track customers' purchasing behaviors and to optimize its product lines and targeted marketing. This statistical approach was so effective that Tesco began applying it to other operations as well and was able to expand its market share by more than 10% over the next decade. Tesco acquired a majority stake in Dunnhumby in 2006.

12 Elizabeth Dwoskin, "Universities Go in Big for Big Data," *Wall Street Journal*, 28 Aug. 2013.

13 James Manyika et al., "Big Data: The Next Frontier for Innovation, Competition, and Productivity." *McKinsey Global Institute*, May 2011.

14 Dana Gardner, "Cloud and Big Data Give Scientists Unprecedented Access to Essential Climate Insights," *ZDNet*, 13 Aug. 12.

15 George Leopold, "Google Invests $130 Million in Cancer-Fighting Big Data Firm," *datanami*, 21 May 2014.

16 Bruce Upbin, "Monsanto Buys Climate Corp For $930 Million," *Forbes*, 2 Oct. 2013.

17 Joao Medeiros, "Jake Porway Wants to Turn His Network of Scientists into a League of Information Champions," *Wired*, 4 June 13.

18 "FAQs," *United Nations Global Pulse*, <http://www.unglobalpulse.org/about/faqs> (15 Aug. 2014).

19 Lucy Bernholz, "How Big Data Will Change the Face of Philanthropy," *Wall Street Journal*, 15 Dec. 2013.

20 Mayur Patel et al., "The Emergence of Civic Tech: Investments in a Growing Field," *Knight Foundation*, Dec. 2013.

21 Kenneth Cukier and Viktor Mayer-Schönberger, *Big Data: A Revolution That Will Transform How We Live, And Think* (New York: Eamon Dolan/Houghton Mifflin Harcourt, 2013).

22 Graeme Wood, "Anthropology Inc.," *The Atlantic*, March 2013.

23 George Orwell, "Politics and the English Language," *Horizon*, April 1946.

24 Pedro Less Andrade et al., "From Big Data to Big Social and Economic Opportunities: Which Policies Will Lead to Leveraging Data-Driven Innovation's Potential?" in *The Global Information Technology Report 2014: Rewards and Risks of Big Data*, INSEAD, Cornell University and the World Economic Forum, 24 April 2014.

CHAPTER
3

ABOUT THE AUTHOR

John Raidt is vice president of Jones Group International, a Northern Virginia-based consultancy. A former staff director of the U.S. Senate Committee on Commerce, Science, and Transportation, he is a fellow with the Atlantic Council, a scholar with the U.S. Chamber of Commerce Foundation, and the author of *American Competitiveness.* Prior to his service as deputy to General James Jones, Special Envoy for Middle East Regional Security, he served as a staff member of key national commissions including the 9/11 Commission, the Commission on the National Guard and Reserves, and the Independent Commission on the Security Forces of Iraq.

THE COMPETIVENESS AGENDA

BY JOHN RAIDT

Key Takeaways

A vibrant and dynamic STEM workforce is critical to the competitiveness of a data-driven economy.

Without access to vibrant and dynamic broadband, it is not possible to realize the full potential of a data-driven economy.

Realign publicly funded R&D to develop data-driven innovation capabilities as well as public-private collaboration in data sharing.

Trade agreements and practices need to ensure the flow, use, and proper protection of data.

A national strategic plan for properly aligning public policies, resources, and priorities is needed to facilitate the beneficial development of a data-driven economy.

In "The Adventure of the Copper Beeches," Sherlock Holmes exclaims, "Data! Data! Data!…I can't make bricks without clay."[1]

Big Data is enabling the U.S. business system to produce revolutionary bricks from exotic new clay that can transform the nation's global economic competitiveness; that is, America's capacity to stimulate investment and create high-quality jobs.

The opportunity could not be timelier. While the forces of globalization are opening vast new consumer markets to our goods, services, and solutions, we face intensifying international competition to be the supplier of choice. Warns the Council on Competitiveness, America must "innovate or abdicate" its global economic leadership.[2]

A High-Stakes Competitive Playing Field

Though America did not discover data-driven innovation—a phenomenon as old as civilization itself—we have set the standard of excellence by applying the indispensable catalysts: incentive, entrepreneurship, and freedom. A new dimension with vast possibilities, however, is fast materializing thanks to a confluence of trends—spurred by technologies pioneered mostly by American firms—that enable us to tap massive digital data flows.

From this accumulating wealth of digital clay, generated by the proliferation of information and communications technology, we are able to extract fresh insights, create novel capabilities, and shape new industries. These competitive assets can catapult the American economy to new heights of leadership and prosperity.

As Moore's law continues to assert itself across the spectrum of information and communication technology (ICT)—from which the Internet of Everything (IoE) is emerging—data production and the ability to process it faster and more affordably will multiply at an Olympic pace. In a *McKinsey Quarterly* article, "The Second Economy," author Brian Arthur observes that "with the coming of the Industrial Revolution…the economy developed a muscular system in the form of machine power. Now it is developing a neural system."[3]

What he means is that in the same way that the human body transmits information to the brain, enabling high-order human function, our industrial, economic, social, and environmental limbs

now have the capacity to produce a rich stream of digital information such that manmade systems can function more cognitively and productively.

GE estimates the gain to global GDP from this advanced system could top $15 trillion by 2030,[4] and Cisco notes that the IoE could yield a 21% growth in global corporate profits.[5] *CIO* reports that "if the cost savings and efficiency gains from the industrial Internet can boost U.S. productivity growth by 1 to 1.5 percentage points, the benefit in terms of economic growth could be substantial, potentially translating to a gain of 25% to 40% of current per capita GDP."[6]

For an economy that has been idling in sub-2% GDP growth, bedeviled by stubbornly high unemployment, stagnated wages, and massive fiscal imbalances, the emergence of a powerful new catalyst for growth is providential. This is particularly true considering that the pull of capital and business operations toward growth markets abroad (mainly in Asia) is so forceful. Mastering data-driven innovation sectors, such as analytics, will enable U.S. firms to deliver a host of new high-value services to global customers via the Internet, while also harnessing the know-how we derive to compete better in every commercial field.[7]

At the enterprise level, data-driven innovation can propel businesses to new levels of productivity, profitability, and competitiveness. In a survey conducted by Oxford University and IBM, "63% of respondents reported that the use of information—including Big Data and analytics—is creating a competitive advantage for their organizations."[8]

Those who fail to grasp the importance of this new reality are at risk. McKinsey reports, "Across sectors, we expect to see value accruing to leading users of big data at the expense of laggards, a trend for which the emerging evidence is growing stronger."[9] *The Economist* puts it more starkly: "Companies that can harness big data will trample data-incompetents."[10]

Yet, success is not self-generating. American firms must compete vigorously on the basis of capability, quality, price, and accessibility for worldwide market share in every product and service category. To compete at our best, U.S. excellence in harnessing Big Data to drive innovation across four domains is a national economic imperative.

Direct Innovation: Big Data is a major industry in and of itself. The United States has an enormous competitive stake in being the leading developer of goods and services across the Big Data supply chain.

Derivative Innovation: Big Data can be harnessed to produce big insight that will enable our economy to produce better goods, services, and solutions across the commercial spectrum.

Enterprise Management Innovation: Big Data can be employed to invent better business models and decision-making processes that make enterprises more successful.

Systemic Innovation: Big Data and data-driven innovation can be used to improve national economic policy and optimize the variables that produce a fertile business environment.

To best understand the advantages, opportunities, challenges, and imperatives in the Big Data age, each of these domains is discussed in depth.

Direct Innovation

In many respects, the Big Data industry is akin to energy development or mining—a broad and integrative process of producing and refining a raw material to power a broader set of economic activities. Here we examine some business components in the main links of the Big Data supply chain, each of which bears significantly on our competitive fortunes.

Generation and Collection

The number of people and institutions using the Internet to socialize, communicate, and shop has accumulated rapidly alongside the proliferation of communication devices, giving rise to an explosion of mobile access, online services, and electronic transactions. The amount of data being created and collected by these forces, already growing 40% per year, will continue to swell at exponential rates.[11] As a result, the race to market cheaper, faster, and better ICT equipment and services, as well as accommodate the gargantuan data flow they produce, will continue to be a hotly contested economic battle space.

Among the fastest growing sources of data generation is sensor technology. Cheap, highly capable sensors can be embedded in

almost anything to transmit electronic data via inexpensive wireless links. The technology is proliferating at a rate of nearly 30% annually, ushering in an era of sensor ubiquity—in the environment, machines, networks, and even in the human body.[12]

BCC estimates that the worldwide sensors market was $56.3 billion in 2010, growing to $91.5 billion by 2016 and $154 billion by 2020.[13] In an article on the "Smart/Intelligent Sensor Markets," MarketsandMarkets estimates that "the total revenue of global smart sensor market is expected to grow at an estimated CAGR of 36.25% from 2013 to 2020."[14]

Joining sensors in revolutionizing data generation is robotics technology. Robots are capable of streaming enormous amounts of data about the functions they perform, measure, and monitor. This is a growing field with large economic potential. BCC Research reports that the global robotics industry, worth $17.3 billion in 2008, should top $22 billion this year and exceed $29 billion by 2018.[15] Amazon reports that the use of robotics could bring efficiencies, saving the company up to $900 million per year.[16] The race for leadership in supplying and servicing this high-tech, high-wage field is moving into high gear, as will the competition to reap the data-driven gains that robotic data streaming can generate.

Transmission, Storage, and Security

Regardless of how digital information is generated, without the ability to electronically transmit, store, and access it, little value can be extracted. For this reason, improving the ability of wired and wireless broadband technology to transmit electronic data is a commercial opportunity and competitive necessity for the Big Data era to fully blossom.

As the pioneer of the Internet, the U.S. innovation system will be relied upon for the products and solutions that can keep the information superhighway untangled and operating at peak efficiency, which allows bigger datasets to move with the speed and reliability required in today's on-demand economy. The global competition across ICT markets will remain intense, particularly as the demand for computers, mobile phones, Internet access, and broadband connections expands in rapidly growing but price-sensitive consumer markets abroad.

There is also an enormous competitive advantage for innovators in the business of storing massive amounts of digital information so it can be aggregated, retained, and organized for analysis and use. One of the larger breakthroughs in data storage has been the advent of cloud computing. Earlier this year, *Forbes* reported that end-user spending on cloud services could exceed $180 billion by 2015, with a global market for cloud equipment climbing to $79.1 billion by 2018.[17]

Accessing this vast potential, however, demands information security. The Big Data era confers big responsibilities to ensure data security and integrity. Without proper security, Big Data (and by extension, data-driven innovation) will flag as data breaches and misuse undermine essential public support. Tremendous comparative advantage will be enjoyed by nations whose business systems, infrastructure, and polices can provide the best security for sensitive data. Moreover, the competitive opportunity for U.S. firms to fill growing demand for data protection solutions (likely to be spurred by national and international policies, laws, and regulations) is enormous.

Analytics

Earlier, we likened Big Data to a raw material, the true value of which is created in processing the resource into useful products. The refining of Big Data into insight, knowledge, and actionable information is the job of analytics. An article in *The American,* "The Next Great Growth Cycle," notes that "Big Data analytics and services, non-existent just a few years ago, is already a $3 billion industry and will be $20 billion in a half-decade."[18]

The all-too-common misperception is that Big Data analytics is about studying consumer behavior to improve marketing effectiveness. Though valuable to producers and consumers, this purpose is merely the surface of a far richer and deeper enterprise of creating data-driven insight in every domain. These include: solving scientific mysteries; identifying meaningful patterns and anomalies; discovering important connections, correlations, and causes and effects; measuring outcomes; shedding light on the dynamics and performance of complex manmade and natural systems; and helping machines learn and perform more capably. For instance:

Drug companies sharing anonymized clinical trial data are producing fresh insight into pharmaceutical safety and effectiveness. Massive amounts of genomic and clinical data are informing the development of better pharmaceuticals and therapies and opening promising new doors in cancer research.

Environmental sensors that measure complex variables are enabling data-based precision agriculture to increase farm yields, helping utilities manage energy demand and improving weather forecasting to support smoother running delivery schedules and supply chains.

Machine-to-machine data flows help monitor systems (e.g., jet engines), providing operational insight, pointing to preventative maintenance requirements, and improving safety and efficiency.

Extracting knowledge from data to inform innovation and business decisions across the economy is big business. The possibilities are astronomical, bounded only by the limits of our capacity to create insight-producing software, algorithms, and computer models.

Distributed Innovation

The book *Moneyball* chronicles the exploits of Oakland A's President Billy Beane, who used data to pioneer cost-effective, winning baseball.[19] It was not the data that was novel but rather Beane's ability to extract meaning and cleverly employ it.

As a result, the A's (a small-market team) achieved an average of 96 wins per season at an average cost of $2 million per player. Compare that to the New York Yankees' average of 99 wins per season, which costs some $5.8 million per player.[20] Using our national pastime to model the possibilities in Big Data is appropriate but equally microscopic compared to the profound ramifications across all industries and activities. Here we take a deeper dive into some of the areas in which data-driven innovation can make U.S. industries and firms more competitive.

Retail and Marketing

Search engines, websites, electronic transactions, and indeed, all of the 21st century data-making tools and activities generate cascades of information that data scientists can mine to understand the demographics, needs, wants, and preferences of entire markets. This enables marketers to better understand customer motives and behavior and also makes possible tailored, intelligent marketing and customized manufacturing (through data-enabled smart machines and 3D printing) that is hastening the decline of the one-size-fits-all era for many products and services.

The McKinsey Global Institute reports that global personal data service providers generate $100 billion in revenue each year, and retailers who use these services are enjoying a 60% increase in net margins.[21] How is such data used to grow revenue and efficiency? Take, for example, Amazon, which uses a recommendation engine to promote products to targeted consumers. According to MGI, some 30% of Amazon's sales are generated by this recommendation engine.[22] Or take the example of XO Communications: after identifying factors that can suggest a customer will depart, the company improved its customer retention rate by 26%. This translated into an annual net gain of $3.8 million.[23]

Manufacturing

The American magazine reports that computational manufacturing "is poised to become a trillion dollar industry, unleashing as big a change in how we make things as did mass production in an earlier era, and as did the agricultural revolution in how we grew things. It is a manufacturing paradigm defined not by cheap labor, but high talent."[24]

Electronic sensors embedded in machines, from consumer goods to factory equipment, stream data to both producers and users about the equipment's design, performance, and maintenance. Sensory data (such as vibration, pressure and voltage) can be used to improve the operational efficiency and the productivity of product design and manufacturing processes.

For example, data generated by heavy machinery (such as aircraft engines or power plant components) can offer insight into operation, helping ensure that systems are operating at maximum efficiency, which cuts the cost of energy and other inputs. GE reports, "In the commercial aviation industry alone, a 1% improvement in fuel savings would yield a savings of $30 billion over 15 years. Likewise, a 1% efficiency improvement in the global gas-fired power plant fleet could yield a $66 billion savings in fuel consumption."[25]

McKinsey cites the use of data-driven innovation in helping achieve a "50% decrease in product development and assembly cost for manufacturers."[26] The ability for Big Data to make our products better and cheaper will enable U.S. manufacturers to compete far more successfully in highly price-competitive global markets, where we must vie against competitors from countries with far lower labor and operational costs.

The downstream economic and competitive implications of data-driven productivity are profound. The global demand for the sensing equipment and analytical algorithms needed to produce industrial efficiency is enormous. Further, consider the benefits of lower transportation, shipping, and utility costs on U.S. industrial and residential customers who enjoy a significant portion of the surplus such productivity generates, freeing up resources for use elsewhere in the economy.

Simafore Analytics identified seven areas in which Big Data and the era of sensor and software-based operations and machine learning is transforming manufacturing.[27] These include:

Engineering Design: Using historical data to select optimal engineering "parameters, actions, and components."

Manufacturing Systems: Using data-based "machine learning and computational intelligence" for better control of manufacturing systems.

Decision Support Systems: Using data-based tools like Neural On-Line Analytical Processing System to coordinate production processes.

Shop Floor Control and Layout: Using "knowledge generated from mining historical work in process data" to optimize floor control and layout.

Fault Detection and Quality Improvement: Using Big Data for success, defect, and failure pattern identification.

Preventative Maintenance: Using historical data and predictive analytics to maintain systems.

Customer Relationship Management: Using customer demand data to modify product design features to meet the customer's needs.

Healthcare and Wellness

The healthcare sector generates colossal amounts of data from research and patient care. The agglomeration of digitized genomic and clinical information, together with the proliferation of biosensors and the growth of e-health records and telemedicine, will add substantially to this data flow. This accreting mass of information contains insight from which medical researchers can produce life- and cost-saving, preventative and therapeutic patient care, as well as more efficient healthcare administration.

Dan Foran, head of informatics at the Rutgers Cancer Institute of New Jersey, told *Scientific American*, "When you go see a physician…you're relying on his past experience. What we're doing now is training the computer to look at large cohorts of thousands and hundreds of thousands (of patient data). It's as if the doctor were making treatment decisions based on the personal experience of hundreds of thousands of patients."[28]

Beyond better healthcare, there is a matter of cost. Healthcare costs in the United States dwarf those of our competitors and continue to grow at a much faster pace than abroad. Yet, as with other industries, Big Data can be used to find efficiencies and cost savings in how medical products and care are delivered. As such, the competitiveness implications of such savings are staggering. In harnessing the power of Big Data, McKinsey estimates $300 billion in potential value to U.S. healthcare alongside an 8% reduction in costs.[29]

The sharing by drug companies of anonymized clinical trial data is producing fresh insight into pharmaceutical safety and effectiveness. As algorithmic and crowd-sourced analysis of medical trials produce better medicine, the cost of expensive litigation and liability judgments will ease. What is more, data-driven innovation can be applied to the business of healthcare delivery. CIO reports that "a 1% reduction in processing inefficiencies in the global healthcare industry could yield more than $63 billion in healthcare savings."[30]

Workforce Development and Industry Impact

There are cascading benefits that can be realized across numerous industries. In the same way that exploiting Big Data can yield insight to improve healthcare systems, it can be used to generate competitiveness-enhancing improvements in

human capital development. Nothing is as vital to U.S. competitiveness and economic success as a highly skilled workforce able to meet the requirements of a high-tech economy where a mastery of science, technology, engineering and math (STEM) skills is paramount.

Big Data can be used to understand how people learn and the factors that lead to student failure. This will enable experts to devise more effective teaching and training techniques customized to individuals and micro-segments based on how they learn best. Other industries that can capitalize on the fruits of Big Data include:

Insurance: The development and use of predictive algorithms enable insurance companies to better measure and price risk.[31]

Food and Agriculture: Optimizing inputs to improve farm yield, prevent food waste, and better manage supply chain and perishable inventory.

Transportation: Improving design of aircraft, locomotives, and automobiles, managing the maintenance and use of fleets, and identifying faster, cheaper modal and routing decisions.

Energy and Infrastructure: Designing more functional facilities, materials, and production methods while improving efficiency and reliability across systems.

Enterprise Management Innovation

A third domain in which data is reshaping America's competitive landscape is enterprise organization, management, and decision making. Companies involved in high-volume transactions and that operate large databases are exploring how to shift their business model, strategy, and value proposition to capitalize on the marketability of their data.[32] To take full advantage, successful companies are establishing new executive positions, such as Data Officer, Analytics Officer, and Data Scientist.[33]

With Big Data, enterprises can (with greater speed and accuracy) better manage inventories, assets, and logistics, as well as set optimal prices based on the most up-to-date market information. Among the most powerful tools being created are algorithms and predictive models using data to provide foreknowledge. As an example, "one global beverage company integrates daily weather forecast data from an outside partner into its demand and inventory planning process."[34]

Advanced analytics provide a basis for swift, trustworthy, fact-based decision making, enabling enterprises to stay ahead of high-velocity change in markets and the competitive playing field. By collecting data from their business units, enterprises can develop dashboards to monitor and better understand systemic and organizational performance, which can help drive productivity, quality, and profitability. In some cases, data-driven innovation is able to remove the mistake-prone human element through smart systems that auto-decide or self-adjust operations.

Erik Brynjolfsson, director of the MIT Center for Digital Business and a top expert on the effect of IT on productivity reports, notes "a shift from using intuition toward using data and analytics in making decisions…Specifically, a one-standard-deviation increase toward data and analytics was correlated with about a 5% to 6% percent improvement in productivity and a slightly larger increase in profitability in those same firms."[35]

Business Environment Innovation

As investors, corporate planners, and entrepreneurs make decisions about where to deploy capital, establish business operations, and create jobs, they look carefully at the quality of the business environment they are considering. In today's global economy, they have many choices as nations compete to offer the most desirable business environment. The factors that make up an attractive business environment include: access to ample customers; reasonable costs; affordable finance; a highly skilled quality workforce; world-class energy and infrastructure; a sound fiscal and monetary system; good governance; and a fertile innovation system.

Big Data and data-driven innovation can be employed to improve America's performance in each category through what CAP calls an "empirical approach to government."[36] Bill Bratton, the well-known big city police chief, pioneered the use of Big Data to chart the location and circumstances of violent crime. The insight enabled preventative strategies that made the community safer and businesses more secure.

Big Data analytics can yield insights that lead to a better understanding of how the economy

functions and the likely effects of laws, taxes, regulations, and policies on society and the economy—critically, *before* they are enacted. Data can also be used to generate more accurate economic data and the effects of economic, fiscal, monetary, and regulatory policies. This comes in addition to: preventing and mitigating threats to public health and national security; improving education systems; offering more efficient public services; and ferreting out fraud, waste, and abuse.

The U.S. Comparative Advantage

The United States is better positioned than any other country to lead and gain a first-mover advantage in the data-driven revolution. It was American innovators who pioneered ICT, creating the Internet, advanced computer science, personal computing, and mobile phones. We remain at the forefront in data-driven architecture, which includes sensor technology, robotics, analytical software, electromagnetic spectrum efficiency, and nanotechnology, to name a few.

More than this historical leadership, America's innovation system is the world's most fertile. The country's firms and institutions hold more patents than those in any other country. The United States continues to hold a decisive global qualitative technological edge. We have the best national labs, technology clusters, innovation hubs, and research institutions. The majority of the world's top universities are located in the United States.[37] No surprise then that the world's 10 most innovative companies in Big Data are located here,[38] and some 90% of the top 500 supercomputing systems used around the world are made by U.S. companies.[39]

In their article, "The Coming Tech-led Boom," authors Mark Mills and Julio Ottino observe that "we sit again on the cusp of three grand technological transformations with the potential to rival that of the past century. All find their epicenters in America: big data, smart manufacturing, and the wireless revolution."[40]

Among our many advantages, the United States has tremendous cultural and professional diversity, possessing unique perspectives and unparalleled expertise across the sciences and in cross-disciplines, where perhaps the richest analytical discoveries will be found. We offer political stability (relative to other nations), opportunity, and a high quality of life that still attracts the world's best minds. We enjoy wide freedoms backed up by law, including the liberty to collect, analyze, and use information. We value and foster entrepreneurship. And as good as we are at competing, Americans are equally keen on collaboration, which is a necessity for world-class innovation.

Data Obstacles And What Others Are Doing To Catch Up

Despite America's inherent advantages in data-driven innovation, we must overcome a number of pitfalls, threats, and obstacles to excel in the face of international competition. These include:

Human Capital

One of the reasons for the enormous shortfall detected by McKinsey in the analytical skills of our workforce is that too few of our people are studying and specializing in the STEM disciplines so critical to the Big Data industry.[41] According to standardized international testing, our student body is performing woefully in STEM compared to their peers in other countries.[42]

R&D Investment

We continue to lag behind other nations in public spending on R&D as a percentage of GDP. The lion's share of federal research funding goes to the life sciences at the expense of critical data-driven initiatives, such as high-performance computing, data modeling, simulation, and analytics. Private sector research is constrained by short-term pressure to meet earnings targets rather than investment in long-term competitiveness.

Fear and Unawareness

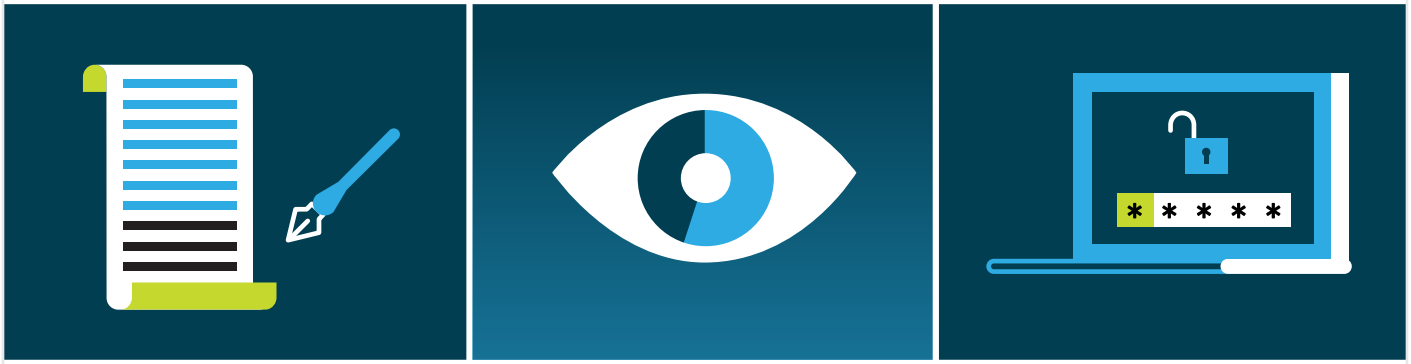
Big Data's potential to improve life can be impeded or even derailed by public apprehensions about privacy, job loss, official ignorance about the opportunities, and/or government mismanagement. We need an enlightened national dialogue on data and innovation to foster a well-informed public and officialdom about the opportunities, stakes, risks, and requirements involved.

Rules of the Road

Every great economic transformation requires modern "rules of the road" to reconcile conflicting interests. There are persistent questions about who owns, secures, and can access data. Industry, consulting with customers and the public, must adopt proper codes of conduct, best practices, and ethical guidelines dealing with data ownership,

PRIVACY: BIG DATA'S BIGGEST OBSTACLE

CONCERNS ABOUT THE MISUSE OF PRIVATE (AND ESPECIALLY SOCIAL) DIGITAL DATA DOMINATE THE BIG DATA DISCUSSION, OVERSHADOWING EVERYDAY EXAMPLES OF DATA-DRIVEN INNOVATION.



68%

of Internet users believe current laws are not good enough in protecting people's privacy online

55%

of Internet users have attempted to avoid observation by specific people, organizations, or governments

21%

of Internet users have had an email or social media account compromised or hijacked

**Pew Research. Anonymity, Privacy, and Security Online. September 2013*

access, and use to establish trust and legitimacy. Lawmakers and regulators need to be prudent and well-informed to get the rules right.

Privacy and Civil Liberty

The misuse of private data, the breach of personal medical and financial information, and the potential for data-based profiling that might violate individual rights and opportunities are legitimate concerns. Without public trust, the enormous good in Big Data cannot be brought to fruition. Yet, perhaps ironically, it is data-driven innovation that can enhance the technical and procedural means for protecting privacy and civil liberties. By fostering public trust and showing the world how it's done, the U.S. business system has the opportunity to make privacy protection a comparative advantage, rather than an impediment to innovation.

Cybersecurity

Part and parcel of data protection is cybersecurity. The Center for Strategic and International Studies says bluntly that the United States is unprepared to defend its computers and networks against myriad cyber threats.[43] McAfee underscores this, reporting that "if there is a race among governments to harden their civilian infrastructure against cyber-attack … Europe and the United States are falling behind Asia."[44] Cybersecurity will undoubtedly continue to play a role in how much trust the public places in the Big Data revolution or the faith that Big Data companies place in the United States.

Infrastructure

The growth in mobile commerce and the broader use of electromagnetic spectrum for wireless communications will be equally explosive. Improved capacity and efficiency of our infrastructure (such as wired and wireless

broadband networks) and the U.S. electric grid must keep pace. No element of the Big Data ecosystem can function in the absence of a reliable and affordable supply of electricity. Without energy, electronic 0s and 1s can't exist, much less tell their story. Thus, there's no overstating the national and corporate competitive advantage in having ample energy delivered on demand via a world-class electric grid. Nor can one overstress the benefits of being the first mover in providing solutions to every need related to the efficient electric generation, distribution, and usage, such as "smart grid" technologies. Despite massive needs, current U.S. infrastructure spending is about the same as it was in 1968, when the country's economy was much smaller.[45] According to the World Economic Forum, the United States ranks 33rd worldwide in "quality of electricity supply."[46] In 2012, we ranked 17th in the UN International Telecommunication Union ITC development index.[47]

Trade and International Rules

For the United States to take full competitive advantage of data-driven innovation, we need access to international information flows and to markets abroad for our services. A Progressive Economy report on 21st-century trade policy notes that "no international agreement protects the free flow of data across borders in the way that the GATT system has provided for the free flow of goods."[48] Coherent national and international norms and rules on data flow, cybersecurity, privacy, trade in services, and IP rights (which are easier to steal in the digital world) are essential. Governments around the world are attempting to gain greater control over the flow of information to serve political objectives. These efforts take many forms, including: capricious standards and regulations on content, data sharing, and Internet access; arbitrary stipulations on the location of servers and data storage facilities; and anti-competitive controls on the information technology supply chain. Firewalls and disparate national rules governing the Internet will turn the global information superhighway into a balkanized collection of back alleys and barricaded side streets impeding mankind's progress in harnessing Big Data for good.

Excess and Irresponsibility

Despite the great potential in Big Data, it must be approached with a sense of humility and deep responsibility. The "garbage in/garbage out" rule can mean that inaccurate, unrepresentative, or improperly analyzed data can result in big mistakes and giant failures. As the *Harvard Business Review* notes, no matter how comprehensive or well-analyzed, Big Data needs to be complemented by "big judgment."[49]

Conclusions

Each of the areas discussed above are hurdles to be overcome but also competitive opportunities to meet global needs with our world-leading strategies, policies, practices, technologies, and services. Other countries are embracing Big Data and building strategies to seize the economic high ground. The European Union has embraced the IoE, undertaking an extensive Big Data Public

"FIREWALLS AND DISPARATE NATIONAL RULES GOVERNING THE INTERNET WILL TURN THE GLOBAL INFORMATION SUPERHIGHWAY INTO A BALKANIZED COLLECTION OF BACK ALLEYS & BARRICADED SIDE STREETS IMPEDING MANKIND'S PROGRESS IN HARNESSING BIG DATA FOR GOOD."

Private Forum and developing strategic research and innovation priorities to capitalize on data-driven innovation.

Among a nation's greatest assets for seizing the Big Data future is supercomputing. While this is an area the United States has long led, Europe, Japan, India, and others are investing heavily to catch up—and with the eye to surpass us. They are gaining ground.[50] Not only is Europe working to build better supercomputers, the EU is also striving to provide high-power computing support for small and middle-sized businesses. These are important competitive developments. What's clear is that leadership in supercomputer infrastructure and access is not some academic luxury but a competitive necessity for the United States.[51]

To be sure, Asia's massive markets are attracting manufacturing, which in turn attracts innovation. China's 800 million (and growing) mobile phones and potential Internet connections dwarf the scale possible in the United States. This is an advantage but one that will be greatly diminished as China erects a "great firewall" on its Internet. A similar innovation-dampening approach is being seen in India, *Progressive Economy* reports, as the country is "requiring telecommunication companies to locate their servers in a country where they can be controlled and hand over data."[52]

Developing countries are well-positioned to bypass expensive legacy computer and communication systems and jump directly to the state-of-the-art networks that can support new industries and the latest thinking. Again, this is an advantage but one diluted by economic, infrastructure, and human resource challenges in developing regions.

The reality is that no nation is better positioned, from top to bottom, than the United States to seize Big Data as a conduit for innovation and global economic competitiveness. Helping the world learn from and make use of the globe's accreting mass of data is a huge business America is uniquely capable of leading.

The renowned business guru W. Edwards Deming famously said, "In God we trust. All others bring data." Bring it we must to renew America's competitiveness and lead the way into a promising new epoch of human advancement.

ENDNOTES

1 Arthur Conan Doyle, "The Adventure of the Copper Beeches," *The Strand Magazine,* 1892.

2 "Innovate America: National Innovation Initiative Summit and Report," *Council on Competitiveness,* 2005, 8.

3 W. Brian Arthur, "The Second Economy," *McKinsey Quarterly,* October 2011.

4 Peter C. Evans and Marco Annunziatta, "Industrial Internet: Pushing the Boundaries of Minds and Machines," *GE,* 26 Nov. 2012, 3.

5 Joseph Bradley, Joel Barbier, and Doug Handler, "Embracing the Internet of Everything To Capture Your Share of $14.4 Trillion," *Cisco*, 2013, 3.

6 Thor Olavsrud, "Big Data Will Drive the Industrial Internet," *CIO,* 21 June 2013.

7 Edward Gresser, "21st Century Trade Policy: The Internet and the Next Generation's Global Economy," *Progressive Economy,* 31 Jan. 2014, 1.

8 "Better Business Outcomes with IBM Big Data and Analytics," *IBM Software,* January 2014, 2.

9 James Manyika et al., "Big Data: The Next Frontier for Innovation, Competition, and Productivity" *McKinsey Global Institute,* May 2011, 11.

10 "Building with Big Data," *The Economist,* 26 May 2011.

11 Manyika, "Big Data: The Next Frontier," 11.

12 "Building with Big Data," *The Economist.*

13 Srinivasa Rajaram, "Global Markets and Technologies for Sensors," *BCC Research,* July 2014.

14 MarketsandMarkets, "Smart/Intelligent Sensor Market by Type, Technology, Application and by Geography," *Forecasts & Analysis to 2013-2020,* March 2014.

15 "The Market For Robotics Te chnologies Expected To Surpass $29 Billion By 2018," *BCC Research*, 20 Feb. 2013,

16 Greg Bensinger, "Before the Drones Come, Amazon Lets Loose the Robots," *Wall Street Journal*, 9 Dec. 2013.

17 TJ McCue, "Cloud Computing: United States Businesses Will Spend $13 Billion On It," *Forbes,* 29 Jan. 2014.

18 Mark P. Mills, "The Next Great Growth Cycle," *The American*, 25 Aug. 2012.

19 Michael Lewis, *Moneyball: The Art of Winning an Unfair Game* (W. W. Norton & Company, 2004).

20 Daniel Esty and Reece Rushing, "Governing by the Numbers: The Promise of Data-Driven Policymaking in the Information Age," *Center for American Progress*, April 2007, 5.

21 Manyika, "Big Data: The Next Frontier," 8.

22 Ibid., 67.

23 "Better Business Outcomes," 5.

24 Mills, "The Next Great Growth Cycle."

25 Evans and Annunziatta, "Industrial Internet."

26 Manyika, "Big Data: The Next Frontier," 8.

27 Bala Deshpande, "7 Reasons Why Big Data for Manufacturing Analytics is Yesterday's News," *SimaFore Analytics*, 9 May 2012.

28 Neil Savage, *"Bioinformatics: Big Data Versus the Big C," Scientific American,* 311 no. 1 (2014): S21.

29 Manyika, "Big Data: The Next Frontier," 8.

30 Olavsrud, "Big Data Will Drive the Industrial Internet."

31 "Building with Big Data," *The Economist.*

32 "Better Business Outcomes," 8.

33 Ibid., 3.

34 Brad Brown et al., "Are You Ready for the Era of 'Big Data'?" *McKinsey Quarterly*, October 2011.

35 Erik Brynjolfsson, Jeff Hammerbacher, and Brad Stevens, "Competing Through data: Three Experts Offer Their Game Plans," *McKinsey Quarterly*, October 2011.

36 Esty and Rushing, "Governing by the Numbers."

37 Mark Mills and Julio Ottino, "The Coming Tech-led Boom," *Wall Street Journal,* 30 Jan. 2012.

38 "The World's Top 10 Most Innovative Companies in Big Data," *Fast Company,* 10 Feb. 2014, <http://www.fastcompany.com/most-innovative-companies/2014/industry/big-data> (18 Aug. 2014).

39 Patrick Thibodeau, "China has the Fastest Supercomputer, but the U.S. Still Rules," *Computerworld,* 2 July 2014.

40 Mills and Ottino, "The Coming Tech Led Boom."

41 Manyika, "Big Data: The Next Frontier," 105.

42 "Compete: New Challenges, New Answers," *Council On Competitiveness,* November 2008, 5.

43 "US Lacks People, Authorities to Face Cyber Attack," *Associated Press,* 16 Mar. 2011.

44 Stewart Baker and Natalia Filipiak, "In the Dark: Crucial Industries Confront Cyberattacks," *McAfee Second Annual Critical Infrastructure Protection Report,* 2011, 2.

45 According to the CBO's estimates, if historical spending and revenue patterns continue in the future, the highway account of the trust fund would be unable to meet its obligations sometime during FY 2012. Similarly, for the 2011-2021 period, outlays would exceed revenues and interest credited to the fund by about $120 billion. See, Douglas Elmendorf, "Spending and Funding for Highways," *Economic and Budget Issue Brief* (Washington, DC: Congressional Budget Office, 2011), 6.

46 Lucas Kawa, "America's Infrastructure Ranks… 25th In The World," *Business Insider,* 16 Jan. 2013.

47 "Measuring the Information Society," *International Telecommunication Union,* 2013, 54.

48 GATT is the acronym for the General Agreement on Tariffs and Trade, an international agreement that seeks to reduce tariffs and trade barriers to promote international trade and prosperity. See, Gresser, "21st Century Trade Policy."

49 Andrew McAfee, "Big Data's Biggest Challenge? Convincing People NOT to Trust Their Judgment," *Harvard Business Review,* 9 Dec. 2013.

50 David F. McQueeney, Statement to U.S. House Subcommittee on Technology and Subcommittee on Research, "Next Generation Computing and Big Data Analytics," Joint hearing, April 24, 2013.

51 McQueeney, "Next Generation Computing."

52 Gresser, "21st Century Trade Policy."

CHAPTER

4

ABOUT THE AUTHOR

Dr. Matthew Harding is an economist who conducts Big Data research to answer crucial policy questions in Energy/Environment and Health/Nutrition. He is an assistant professor in the Sanford School of Public Policy at Duke University and a faculty fellow at the Duke Energy Initiative. He aims to understand how individuals make consumption choices in a data-rich environment. He designs and implements large scale field experiments, in collaboration with industry leaders, to measure the consequences of individual choices and the extent to which behavioral nudges and price based mechanisms can be used as cost-effective means of improving individual and social welfare. He was awarded a Ph.D. from the Massachusetts Institute of Technology and an M.Phil. from Oxford University. He was previously on the Stanford University faculty and has published widely in a number of academic journals.

GOOD DATA PUBLIC POLICIES

BY DR. MATTHEW HARDING

Key Takeaways

Big Data is not simply bigger but a deeper layering of many sources of information linked to each other. Data-driven innovation thrives on the ability to link many sources of data into a coherent structure, providing new insights and improving efficiency.

Good corporate and public data policies serve the common good. Big Data needs to be grounded on open standards and requires advanced technological solutions to monitor and enforce high quality in acquisition and use.

Public policies should encourage responsible use of data. Privacy and security concerns are best addressed by industry-led initiatives that are flexible, innovative, and technologically sound. Policies that restrict or prevent data access and sharing are a major impediment to innovation and public welfare.

Big Data is getting bigger every day. It would be more accurate to describe this phenomenon as a data deluge. IT research company Gartner predicts 650% growth rates for enterprise data over the next five years.[1] While scientists do not even agree on when data becomes "big" (since it depends on the relative computational power needed to process it), it is an inescapable fact that it is transforming society at an ever-increasing pace, and it introduces a unique set of challenges and opportunities for today's enterprises.

IBM Chairman, President, and CEO Virginia Rometty argues that "data constitute a vast new natural resource, which promises to be for the 21st century what steam power was for the 18th, electricity for the 19th, and hydrocarbons for the 20th."[2]

This natural resource is not just abundant but also multiplying at astonishing rates. As with all natural resources, we will need to establish a carefully considered system of policies ensuring its productive use, minimizing the extent to which it is wasted or misused, and ensuring that it remains available to entrepreneurs in a vibrant competitive environment. Lastly, we need to carefully consider the potential for damage or unintended consequences resulting from the use of Big Data, since the large-scale use of data may lead to hazards we cannot yet foresee. Ignoring the inherent risks of Big Data and imposing regulatory barriers before human genius has had the opportunity to explore its potential are both damaging to our long-term progress and wellbeing.

To take full advantage of this remarkable new resource, we need to develop responsible public policies that encourage innovation and growth while also protecting individual freedom and advance the common good. If data is going to become a major factor of production, along with capital and labor, public policies will need to create the institutional framework that restricts misuse while promoting healthy competition and protecting the interests of society at large (and the most vulnerable members of society in particular).

At the same time, it is important to acknowledge that just like natural resources, there are many different types of data. Each type has unique features and presents different challenges and opportunities.

Government data from tax records is different from private data from store transactions. Personally identifiable data is different from anonymous data. Some data may cause harm if it becomes publicly available, but other data can and should be accessed by the general public. Heterogeneity is fundamental to the data world, and data varies widely in terms of collection, use, and ultimate impact. When designing data policies, we should be careful not to adopt a one-size-fits-all approach that limits the organic growth of the new and vibrant data economy.

Below, we explore some of the basic building blocks of good data public policy. It is not meant to provide definitive answers, but rather, to highlight some of the main questions that need to be asked and the broader discussions policymakers need to engage.

Depth in Data and Current Policy Principles

The first question we must address is: what is so special about data-driven innovation that it requires us to consider new data policies? Isn't all data the same, and wasn't personal or sensitive data available all along? Tax returns have contained a lot of sensitive data since long before computers turned them into streams of 1s and 0s. Since much of the Big Data discussion today is focused on issues of privacy, it may seem at first glance that existing policies simply need to be "scaled up" to take into account the larger datasets available. But this would be misleading. As we shall see, existing policies are too restrictive to stimulate the innovative potential of data.

Big Data is not simply a bigger version of the data already available. In fact, a better term for Big Data is Deep Data. Data achieves its depth by the layering of many sources of information and allowing these layers to be linked to individuals and between individuals. We use technology to interact with the world around us, and each time we do so, we create a new layer of data. Taken separately, each of these layers of data is of limited use. Together, however, they become a formidable resource that allows an outside observer to understand the motivations and choices of individuals with increasing accuracy. Once a need is identified, the opportunity exists for an innovative entrepreneur to offer a product that is specifically tailored to the individual. Big (or Deep) Data becomes a precisely quantified imprint of all our lives.

This is tremendously exciting, as it promises to revolutionize our lives, but it could just as easily be misused or abused in the absence of well-thought out public policies. The entrepreneurial spirit thrives in a free environment; that is, as long as sound policies promote responsible innovation while safeguarding against practices that are morally repulsive or harm others.

Existing public policies addressing the use of data in society date back to the early 1970s, when the U.S. Department of Health, Education, and Welfare issued "Records, Computers, and the Rights of Citizens," a report outlining safeguards known as the "Fair Information Practice Principles," which formed the bedrock of modern data policies.[3] These principles have guided much of the subsequent legislative activity from the 1974 Privacy Act to the 1996 Health Insurance Portability and Accountability Act (HIPAA). While some of the principles outlined at the time remain valid today, it is time to re-evaluate them in the face of modern advances in data science and the potential of Big Data to promote the public good.

In 2012, the Obama Administration issued a report outlining a proposed Consumer Privacy Bill of Rights that addresses commercial (and not public

"WE USE TECHNOLOGY TO INTERACT WITH THE WORLD AROUND US, AND EACH TIME WE DO SO, WE CREATE A NEW LAYER OF DATA."

sector) uses of personal data and substantially expands on the principles first outlined in the 1970s. While privacy protection has long been a major concern for policymakers, recent events have highlighted the dangers of government abuses of Big Data. This has been enormously damaging to public perceptions of the costs and benefits of sharing personal information, since data-driven innovation is now associated with fears of spying or criminal activities. While privacy principles are important, we also need to realize that we do not have to choose between privacy and prosperity. Thus, it is important to re-evaluate privacy principles in light of today's needs and opportunities and recall that concerns surrounding government use of data and access will not be resolved by imposing new burdensome restrictions on the legitimate use of data by businesses.

When discussing good data policies, we need to first address data policies and best practices that are beneficial to both public and private entities. Policies for data acquisition are an example of this, as the success of data-driven innovation depends on the quality of the raw input (i.e., the data). At the same time, we need to explore the tension between the value placed by society on privacy and the regulators' tendency to impose restrictions on use. We need to evaluate the extent to which existing regulatory frameworks are still suitable in today's data-driven world. These two different areas where good data policies are a necessity are not completely independent of each other either. The propensity for regulatory action diminishes as responsible data acquisition and usage policies are established.

Policies for Good Data Acquisition

Data is acquired from a variety of sources. Whether it is automatically generated by sensors or entered into a spreadsheet by a human being, it is important to think through the process of acquisition, since the quality of the data acquired is crucial for its economic value.

Common Standards

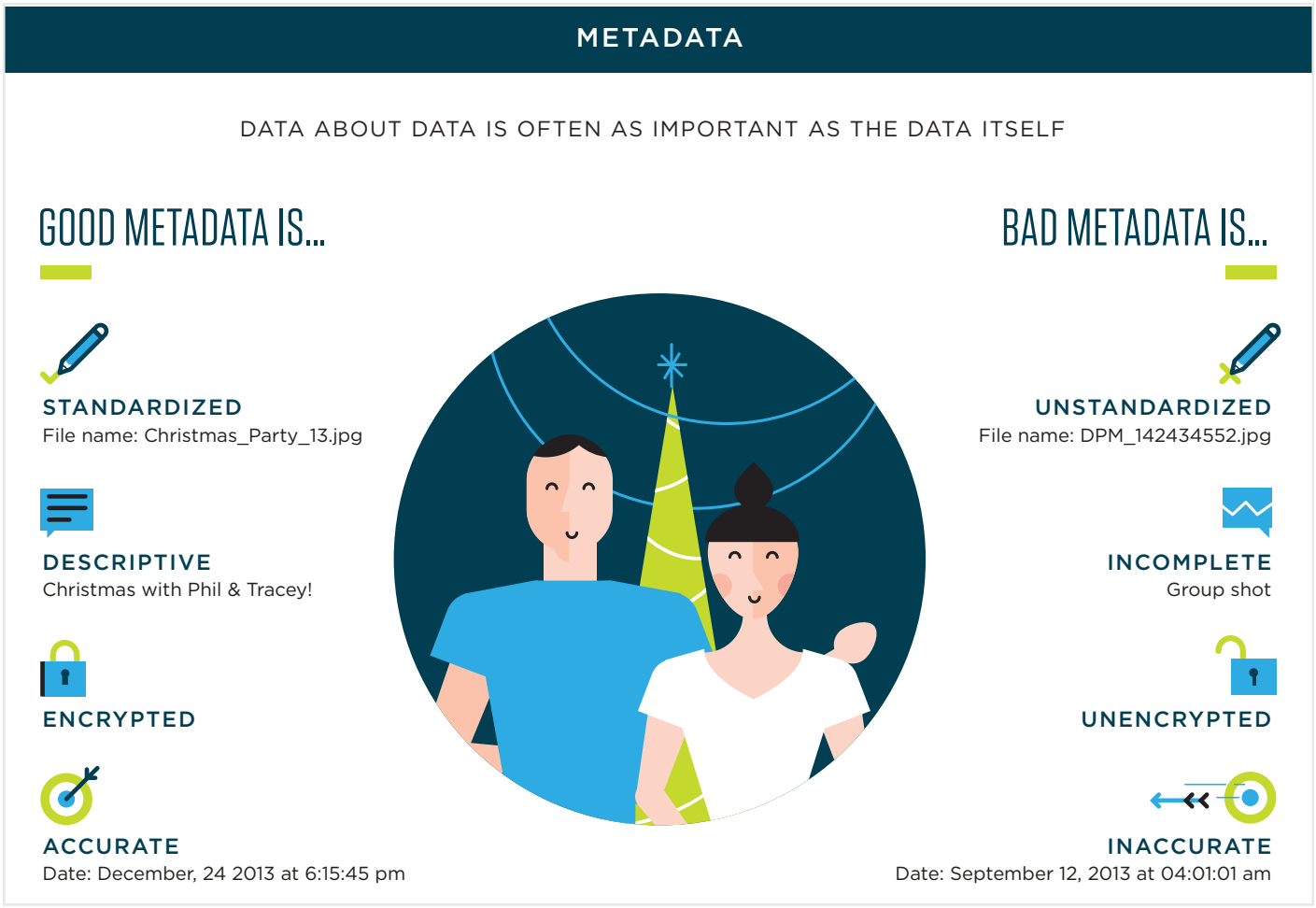
The principle of *data accuracy* calls on any organization creating, maintaining, using, or disseminating data records to assure the accuracy and reliability of the data. We need to further strengthen this principle by ensuring that data is not only collected as accurately as possible but is also subject to *common standards and procedures*. Trade in the early days of the American Republic was limited because colonists brought with them many different units of measurement from England, France, Spain, and Holland. More recently, NASA's Mars Climate Orbiter disintegrated on approaching the Red Planet's atmosphere because the ground computer produced misleading output in non-standard measurement units, which confused the scientists and led to erroneous control commands.

While public datasets have become increasingly easier to access in recent years (particularly with the launch of Open Data initiatives, such as Data.gov), using the data is often confusing and impractical. The innovator looking to access these resources is typically facing a confusing array of data formats, many of which are derived from software packages that no longer exist. It is also common to encounter government records in the form of scanned images, which cannot be easily read by a machine. Policies aimed at establishing the use of *standard open source data formats* are urgently needed to lower the entry barriers to the data economy and facilitate the development of new products and services.

Providing Metadata

In spite of the recent negative publicity associated with the collection of metadata by the NSA, the acquisition and standardization of this type of data needs to be encouraged. Metadata refers to the additional records required to make the raw data useful. It references the information collected about a data entry, which is different from the content itself and includes a description of the source, time, or purpose of the data. For example, it helps clarify the units in which a transaction was recorded and avoids the confusion that arises when the user of the data is not sure if distances are measured in miles or kilometers.

In the popular media, certain types of metadata (like browsing records or the network of phone calls) are well-publicized. While some applications are based on the metadata, in practice, data scientists spend a lot of time cleaning and organizing the metadata in order to be able to make sense of the content of interest. Unfortunately, a large amount of effort is spent in businesses all over the country trying to make sense of both external and internal datasets when metadata is lacking. It is common for the content to be recorded, but little effort is made to document what the data is actually about.



For example, the amount of a transaction may be recorded, but the units are not. Without this additional metadata, we do not know whether the amount refers to dollars or cents. Additional information, such as the time or place of the transaction, would provide a much more detailed picture than a single number. The lack of emphasis on the systematic and standardized accumulation of metadata leads to substantial costs and increases the potential for mistakes. Without metadata, we can easily misinterpret the nature of the data we use and reach misleading business decisions.

When considering how much metadata we ought to collect and associate with a given dataset, it is important to recall one of the fundamental principles of science: *replicability*. When a scientist reaches a conclusion based on the analysis of a phenomenon or experiment, it should be possible for another scientist to reach the same conclusion if she follows precisely the same steps. This ensures that the conclusion is based on fact and not on coincidence.

A similar principle ought to guide the collection of metadata. Irrespective of whether the data is collected for internal or public use, a rich enough set of metadata should be available to allow someone else, at least in principle, to collect the same data again. Occasionally, data collection may involve the use of proprietary technologies or be based on algorithms conferring a competitive advantage to their owner, and open access to the metadata may not be possible. This is likely to be the exception rather than the norm. From a social policy perspective, the benefit of restricting access to metadata may be outweighed by the need for data accuracy.

In the long run, the importance of trust in the marketplace should not be underestimated, and competitive pressures in the private sector are likely to limit the use of data for which no adequate metadata describing its origins and nature is provided. The need for recording metadata as part of a healthy process of data accumulation does not justify its abuse by government agencies. While distinct from the actual content it characterizes,

metadata contains personal information and should be handled with the same amount of care as any other data, since its release may cause harm to the subjects upon whom it is based. As such, encrypting or anonymizing metadata is an important component of safeguarding privacy.

Data Depth and the Value of Historical Data

Data collection policies should also encourage *data depth*. For example, data depreciates at a much slower rate than technology. While a 5-year-old laptop may be obsolete, 5-year-old data may still be very valuable. Since in many circumstances it takes a long time for economic actors to change behavior, many important business or policy questions can only be answered if detailed historical data is available. Unfortunately, it is still common practice for many public and private entities to delete or overwrite their historical data at regular intervals. While this practice made sense when the cost of storage was high, it is difficult to justify today given the plummeting prices for storage and the ubiquitous presence of new storage technologies, such as cloud storage.

To put the dramatic reduction in price into perspective, it is estimated that the average cost of storing 1 gigabyte of data was more than $100,000 in 1985, $0.09 in 2010, and only $0.05 in 2013.[4] Our policies and practices need to keep pace with technological advances in order to make sure we do not miss out on future opportunities. Policies are needed to encourage the preservation of historical data in such a way that it can be linked to subsequent waves of new data to form a more complete image of the world around us. There is an inherent risk in reaching decisions based on data snapshots (no matter how detailed the data content may be) while ignoring the sequence of preceding events.

Addressing Measurement Error

We must also realize that even with the most detailed set of best practices in place, data acquisition is likely to be imperfect and some data will be recorded with error. At the population level, data imperfections themselves are less troublesome if no discernable pattern of bias is present in the data collection. While some biases may be unavoidable, it is important to document them and make users aware of their existence. For example, online data is only representative of the user base for a specific platform and may not be representative for the American population as a whole. Not recognizing this fact may lead to grave decision errors.

Recent policy discussions captured in the 2014 Federal Trade Commission (FTC) report on data brokers recommends that companies that collect personal information create mechanisms that allow consumers to access their information and provide mechanisms for individuals to correct information they believe is erroneous.[5] While this suggestion may play a role in the quality control of highly sensitive personal data, it is unlikely to be of much use in general, given the data deluge we are experiencing today. In practice, it would be impractical for users to engage at the detailed level with every one of the millions of data points generated each day. At the same time, it is important to realize that users can maintain control over the types of data and the uses for which personal data is employed. For example, it is possible for a user to prevent her health information from being shared. Thus, in the aggregate, users can maintain a large degree of control without the need for managing their data every day.

A much more useful policy would encourage data brokers to employ machine learning algorithms to automatically check the validity of the data and tag suspicious entries for further evaluation and potential correction. One of the important aspects of Big Data is the increased speed at which data is generated. Automated systems are an effective and efficient mechanism for validating the accuracy of the data in real time. Policies promoting the automation of these tasks are to be preferred over those that impose additional burdens on consumers. In a world where an ever-increasing number of activities demand our attention, the process of data acquisition needs to remain transparent but unobtrusive, requiring focused interaction with the consumer only when needed. For example, an automated process could detect that my property information is mis-recorded and ask me to correct it. Companies like Opower use property records to provide consumers with tailored energy saving tips, which can help reduce monthly energy bills. A more accurate property record will enable companies like Opower to provide better products and help consumers save money.

Policies for Good Data Use

While many of policies can be implemented in the form of best practices by both the private and public sectors, we should also consider the role that public policy can play in promoting responsible data use and data-driven innovation. Given the ubiquitous presence of data, regulating all sources of data is a quixotic (and economically inefficient) task. The increased penetration of the Internet of Things and the resulting rise in data collection makes it impossible to apply existing principles of data use. Current data use frameworks emphasize the need for consumer consent. As the recent White House report on Big Data highlights, "this framework is increasingly unworkable and ineffective."[6]

The notice and consent approach is outmoded and imposes an unnecessary burden on consumers. It is impossible for us to continuously engage in a process that requires us to agree to extensive notices and for firms to try and anticipate all the possible uses of data in the future. This approach imposes cognitive costs and by its very nature remains incomplete, since future uses of data cannot always be foreseen. In fact, they may not have even been discovered at the point in time when the data is acquired. As a result, the policy prescriptions need to put focus more on principles summarizing our societal agreements on the nature of permissible data applications that are consistent with our values.

The Role of Context

One of the obstacles we face in thinking about policies that will actualize the full value of data to both owners and consumers of data is the outdated emphasis on imposing boundaries on use based on the *initial context* in which data was collected. The idea appears to be that uses of the data should be restricted to the context in which the consumer provided the data. This principle was used historically to determine data use policies and also reappears in the Consumer Privacy Bill of Rights. Leaving aside issues of determining what the context of data generation actually is, it seems an unnecessarily restrictive requirement. Consider the following thought experiment.

If my cell phone uses GPS to track my location in order to provide me with driving directions to the grocery store, does this mean that the location data generated should only be used for providing driving directions? I may choose to use GPS data to provide me with information on better shopping opportunities nearby, inform me about the historical buildings I am driving past, or provide me with tips to improve my driving experience or economize on fuel. Perhaps at some point in the future, the same data can be used to help city planners design better cities or inform businesses about the need to open an additional store closer to my home. The benefits of reusing data are only limited by our imagination, and it is wasteful to limit its use to some "original" context.

The Serendipitous Use of Big Data

Data scientists have recently begun investigating the value of repurposing data for new uses. As seen with the discovery of penicillin, a mix of luck and human ingenuity can spark new data applications. We refer to this process as the *serendipitous use of Big Data*, a process that should be encouraged by public policies and not arbitrarily restricted. As more enterprises can access a variety of data sources, we will see innovative new products and insights emerging. The process of repurposing data is likely to gather further momentum with the increased availability of Open Data (see Chapter 6), and a variety of public and private datasets will be used to challenge established wisdom and will have lasting consequences on society. Repurposing data is not a new process either. In the middle of the 19th century, an entrepreneurial oceanographer and Navy commander repurposed logbooks to determine the best shipping lanes, many of which are still in use today.[7]

We are already seeing many new innovative products created by repurposing data. Examples include:

PriceStats, a company originating in an MIT academic project, collects high-frequency data on product prices around the world and creates daily inflation indexes used by financial institutions.

ZestFinance uses advanced machine learning algorithms combined with numerous data series to create a better risk profile for borrowers and a more precise underwriting process.

Factual combines many different data sources to provide location-based information on more than 65 million businesses and points of interest globally.

Many other products are going to be discovered as we start to make sense of the connections between the available data. Google Correlate is a free tool that allows anyone to find correlations between a time series of interest and Google searches. Online searches have now been shown to be predictive in the short run of many economic phenomena of interest, such as unemployment, housing prices, or epidemics.[8] The exploration of such seemingly arbitrary correlations between datasets can even lead to surprising scientific discoveries. Researchers correlating records of patients with HIV and multiple sclerosis (MS) discovered that the two conditions do not seem to appear jointly, and this might be due to the fact that existing HIV medications are successful at treating or preventing MS.[9] If confirmed, this appears to suggest that treatment for MS may be possible by repurposing HIV treatments. This is a rather stunning example of how correlations between two different datasets can lead to life-changing insights and treatments for patients.

Responsible Use Policies (Rather than Prohibitions)

Given the potential for good resulting from the ability to link many different data sources, we must re-evaluate old data-use principles and ask ourselves what the potential for innovation is and whether we are willing to let it flourish. This is not to say that we are going to avoid *moral dilemmas* along the way or require additional *policies that prevent abuse*.

Data enables more informed decision making by policymakers and empowers consumers to make better choices. Sadly, we often see policymakers deciding to prohibit the use of certain types of data altogether rather than taking a more nuanced approach that allows the public good to flourish. Consider the emotionally charged topic of using data on children and infants.[10] It is certainly true that these are vulnerable populations that cannot consent to the use of their personal data. At the same time, numerous data sources are already available from birth, such as vital records, hospital records, insurance claims, disease registries, and other administrative records. Using these records has enabled researchers to develop many useful insights. For example, access to natality records has given researchers insight into the costs of low birth weight and its large, negative consequences later in life. These types of insights are important and can provide the evidence needed for policies, which are better able to promote the public good. Rather than preventing access, we should be engaging in the deeper conversation of how to allow access and address potential privacy concerns.

Encouraging Responsible Data Use

Public policies promoting responsible data use will need to address valid privacy concerns. As noted above, data comes in many different types. This heterogeneity is essential to the nature of the data-driven industry, and policies need to take this into account. It is not feasible to push the burden of monitoring use onto the consumers by asking them to review and consent to every single use of their data. At the same time, we should resist calls for inflexible top-down regulatory approaches, which fail to distinguish between different types of data or applications. In particular, we ought to be concerned with policies that attempt to block access entirely or require certain types of data records to be destroyed.

Policies like the so-called "right to be forgotten" promoted by the European Union are unlikely to be effective mechanisms for protecting privacy for the vast majority of consumers and may impose unnecessary barriers to innovation. Such an approach is difficult to reconcile with the value we place on freedom of speech and could be

"OUR POLICIES AND PRACTICES NEED TO KEEP PACE WITH TECHNOLOGICAL ADVANCES IN ORDER TO MAKE SURE WE DO NOT MISS OUT ON FUTURE OPPORTUNITIES."

manipulated to create deliberate loss of data with unintended consequences later on in areas such as national security or law enforcement.

We must ask, who is most informed to ascertain the risks and benefits of using sensitive data? Does it really make sense to leave this decision to a remote bureaucracy or trust outdated principles in a data-driven world that is changing so rapidly? Data risks and benefits are best evaluated by the innovators deciding whether to develop a new product or service. They have the most complete information, and we should encourage them to engage in careful reviews of the uses of the data, as well as the potential hazards. This places great responsibility on the industry innovators, since a miscalculation can lead to loss of consumer trust and cause irreparable damage to a company's reputation and profitability. Thus, the creators of new data products have the strongest incentives to address privacy concerns early on. It cannot be stressed enough that evaluating the risks and benefits to consumers of new data-driven products should happen as early as possible in the product lifecycle.

The Importance of Research and Experimentation

Before a product even exists, data is at the core of research activities. The only way to create truly innovative products is to experiment. Rigorous experimentation provides the foundation for uncovering the features of a product that best appeal to customers and deliver the most value. For example, a standard approach in marketing is the principle of A/B testing. This is simply an experiment where customers are presented with either option A or option B of a product. The behavior of the two groups of customers is then observed, and it helps explain which option provides better value for the customers. This option is then offered to the larger population.

Not only do research and experimentation provide important business insights, they also lead to important scientific breakthroughs as we better understand what motivates human behavior. The aforementioned Opower is a new company that uses behavioral nudges to help consumers save on their energy bills. They use Big Data to determine, for each household, a group of other households that are similar in terms of property characteristics or composition. Opower then works with the utility company to present customers with data on how their energy use compares to other households. They also provide targeted energy savings tips. This social comparison has been shown to be an effective low-cost nudge for consumers to become more energy efficient. Opower has refined and also quantified the impact of this approach using more than 100 large-scale randomized experiments involving different messaging approaches on more than 8 million utility customers.[11] Using experimentation, Opower has developed a data-driven product that saved more than 5 terrawatt hours of energy—enough to power New Hampshire for a year.[12]

This type of data-driven experimentation in the real world makes business sense and allows us to develop new and innovative products, which benefits consumers. At the same time, the rigorous randomized controlled trial approach provides us with the scientific rigor needed in evaluating the benefits of such new products. Social scientists have also learned a lot about human behavior and gained insights into people's motivations and perceptions, as well the obstacles they face when trying to adopt good behaviors, such as becoming more energy efficient.

Industry-Driven Solutions for Data Use Risk Certification

In spite of the obvious benefits of experimentation and the use of personal data to develop innovative new products, not all attempts are successful. A recent scientific experiment conducted by Facebook and Cornell University looked at the spread of emotional content in social networks. The experiment has drawn strong criticism in the popular press, despite the fact that it was conducted within the legal scope of existing user data agreements.[13] This opens up the question of what can be done to better address users' privacy concerns and adequately quantify both the risks and rewards involved in the process of data-driven innovation. Scholars have highlighted that the existing framework relying on constant legal notices provides the illusion of privacy at a substantial burden to the consumer.[14]

We need new industry-driven solutions that are flexible enough to promote the responsible use of individual customer data. A promising approach was recently suggested in a *Stanford Law Review* article, which calls for the creation of industry-based Customer Subject Review Boards, loosely modeled on the Institutional Review Boards that

evaluate and approve human subjects-based research in academic institutions.[15]

New Principles for Responsible Data Use
The first step in the process of establishing a credible self-regulatory approach that addresses the privacy and ethical concerns of consumers is a series of discussions involving all stakeholders to establish broad new principles for the responsible use of personal information in the Big Data world. As part of this discussion, we need to reevaluate the current framework on *data ownership*, which is rather vague due to the speed at which the nature of data sharing is changing. In particular, we need to pay attention to the range of possible claims to ownership depending on the source, type, and degree of individual contribution to data generation.[16] There are subtle but important distinctions that need to be addressed between the subject of the data, the creator, funder, or enterprise that collects the data, and the consumer of the data. As part of this broad dialogue involving the different stakeholders, we need to agree upon clear categories of data and how to identify which data are sensitive and thus have the potential for harm if used irresponsibly.

Once new principles of data use are established, a clear process for reviewing new products, services, research, or experimentation using consumer data can be created reflecting these principles. The institutional framework for evaluating whether a given project complies with these principles can vary from business to business to support (rather than slow down) the product development cycle. While some businesses may prefer to create internal mechanisms, others may defer to an outside organization to determine compliance. Over time, a robust system of certification will develop to support this process. The review process will perform a number of important functions and provide the ingredients of a rigorous cost-benefit analysis that is subject to uniform, industry-wide ethical principles established beforehand.

What might such a review involve? The review process will help clarify the exact purpose of the data used in a given product or service. Product developers will be given the opportunity to carefully evaluate the degree to which sensitive data is required and whether Open Data or non-sensitive data alternatives may be easily substitutable.

Once it is established that sensitive data is required, procedures can be put in place to guarantee customer privacy. This may involve technical solutions related to encryption and storage or managerial solutions restricting certain types of data to employees with adequate training and who are essential to a given task. At the same time, it is important to evaluate the inherent risks involved in using personal data. There may be risks, such as emotional distress, to customers from using the product or service. It is also important to consider whether third parties may inappropriately use the product in a way that would be harmful to customers.

Procedures need to be put in place to deal with situations where customers may have additional questions or concerns or may wish to opt-out. Customers will need to be reassured that choosing not to share their data will not be detrimental to them in the future or lead to penalties or loss of benefits. A careful examination of all aspects of data use will help quantify the benefits and risks to the customer and the firm. This process will ensure the responsible use of individual data without the need to impose bans of the types of data or activities that can be explored. This does not mean that all projects will be certified. We might expect that certain projects will be deemed by the review board to be too risky to the consumer or the firm, and a prudent manager will send the project back to the design team to be rethought.

Benefits of Self-Regulation
Carefully reviewing products or services before they are launched will have a positive impact on consumer privacy and will be beneficial to firms in a number of ways:[17]

In today's data economy, consumers receive substantial benefits from sharing personal or sensitive data. Yet, not all firms have a strategy in place for communicating these benefits to consumers. An effective review process will enable firms to engage in a rigorous method of identifying the costs and benefits to consumers of sharing sensitive data. While these may differ from case to case, a clear formulation will make it easier to communicate the benefits directly to consumers.

Managers can anticipate and avoid costly media disasters by better managing the risks involved in developing the product. If a product may expose customers to substantial risk (e.g., if it were

"WE NEED NEW INDUSTRY-DRIVEN SOLUTIONS THAT ARE FLEXIBLE ENOUGH TO PROMOTE THE RESPONSIBLE USE OF INDIVIDUAL CUSTOMER DATA."

hacked by a criminal organization), the review may highlight the need for additional security measures or protocols to reduce the risks resulting from the release of sensitive data.

Increased transparency in the use of sensitive data will assist in addressing regulatory concerns and compliance with existing regulatory regimes. Additional regulations will be preempted by a well-functioning system of addressing privacy concerns while allowing the innovation to flourish.

While this review process is likely to be conducted at the organizational level, demand for certification products is likely to arise in the marketplace. Certification initiatives are likely to develop organically and offer an additional level of certainty that the products have met given minimum risk standards. Certification fulfills a natural role in the marketplace, and while we would not expect there to be demand for certification for every data-driven product, it may help promote common standards and increased transparency in areas where privacy concerns are particularly strong in consumers' minds, such as education or health.

The Need for Supporting Policies

Lastly, it is important to realize that good data policies also require a wide range of additional policies supporting the effort to build a data-focused economy and nurture data-driven innovation. As data becomes more central to our lives, the need for advanced data science skills will continue to increase, and the need for workers with skills in data science and computer programming will become increasingly acute. Policies aimed at teaching these skills in schools will be essential.

A recent *Economist* article points out that in some countries, like Estonia, children as young as six are taught the basics of computer programming.[18] Many countries are already mandating that computer programming be taught in primary schools. While specialized data-science skills will be at the core of tomorrow's job requirements, the need for improved data literacy is already felt at all levels of society. Managers and executives in companies across the country now have a vast amount of data at their disposal and need to learn how best to evaluate the available evidence. Doctors have access to real-time health records from sensors and insights from genetic information; they need to learn how to make data-driven treatment and prevention choices. Consumers can look up detailed information generated by the many communicating appliances in their homes (such as smart thermostats) and develop action plans that help them live healthier and happier lives.

Many of the privacy concerns can be addressed by continuing investment in research to provide new advanced technologies that safeguard sensitive data. Privacy concerns cannot be alleviated by a once-in-a-generation policy rule. Researchers have provided many examples of data once considered secure that, as a result of advances in technology or algorithmic understanding, were later discovered not to be so.[19] This is not a matter of criminal activity or data breaches but rather a result of our constantly improving technologies. Cryptography and anonymization techniques can often be reversed using more advanced algorithms. According to a recent Harvard study, 43% of anonymous data source samples can be re-identified.[20] This does not mean security is unachievable. Rather, robust competition between technology companies is the best approach to developing new security solutions with more effective anonymization techniques.

Today, we have a tremendous opportunity to advance wellbeing by promoting good data public policies that drive innovation through the responsible use of sensitive data. Such policies require best practices to address the use of data throughout its lifecycle, from acquisition to ownership and end use. We must be careful not to be fooled into believing that our only choice is a rigid, top-down set of legislations based on outdated fair use principles that limit innovation

by restricting access to data sources or prevent serendipitous discoveries that come from exploring and combining seemingly unrelated datasets. We need to remain flexible and adapt to the new opportunities that data presents to us and not be afraid to ask the hard questions on what the best approaches are for enabling innovators' access to and responsible use of sensitive personal information.

At the same time, we need to stress the leadership role companies at the core of the data economy have in creating new technological solutions to ensure privacy and security and in forming institutional structures for quantifying the risks and benefits of new data-driven products. If we allow innovation to flourish in a responsible manner, the public good will be promoted by new products and services. We do not face a choice between innovation and privacy but rather between responsible use and a false sense of security that comes from over-regulation and limited access.

ENDNOTES

1 Raymond Paquet, "Technology Trends You Can't Afford to Ignore," *Gartner Webinar*, Jan. 2010.

2 Virginia Rometty, "The Year of the Smarter Enterprise," *The Economist: The World in 2014*.

3 Secretary's Advisory Committee on Automated Personal Data Systems, "Records, Computers, and the Rights of Citizens," *U.S. Department of Health, Education, and Welfare*, July 1973.

4 "Average Cost of Hard Drive Storage," *Statistic Brain* <http://www.statisticbrain.com/average-cost-of-hard-drive-storage> (28 Aug. 2014).

5 "Data Brokers: A Call for Transparency and Accountability," *Federal Trade Commission*, May 2014.

6 Executive Office of the President, President's Council of Advisors on Science and Technology, "Big Data and Privacy: A Technological Perspective," May 2014.

7 Viktor Mayer-Schönberger and Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Houghton Mifflin Harcourt, 2013).

8 Hal R. Varian, "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, 28, no. 2 (2014): 3-28.

9 "HIV and MS: Antithesis, Synthesis," *The Economist*, 9 Aug. 2014.

10 Janet Currie, "'Big Data' Versus 'Big Brother': On the Appropriate Use of Large-scale Data Collections in Pediatrics," *Pediatrics*, 131, no. Supplement 2, 1 April 2013.

11 Hunt Allcott, "Site Selection Bias in Program Evaluation," Working Paper, March 2014.

12 Aaron Tinjum, "We've now saved 5 terawatt-hours," *Opower,* 22 July 2014 <http://blog.opower.com/2014/07/opower-five-terawatt-hour-energy-savings-new-hampshire> (28 Aug. 2014).

13 Adam Kramera, Jamie Guillory, and Jeffrey Hancock, "Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks," *Proceedings of the National Academy of Sciences of the United States of America*, 111, no. 24 (2014).

14 Fred Cate, "The Failure of Fair Information Practice Principles," *Consumer Protection in the Age of the "Information Economy,"* ed. Jane Winn (Ashgate, 2006).

15 Ryan Calo, "Consumer Subject Review Boards: A Thought Experiment," *Stanford Law Review*, 3 Sept. 2013.

16 David Loshin, "Who Owns Data?" *Information Management*, 1 March 2003.

17 Calo, "Consumer Subject Review Boards."

18 "Coding in schools", *The Economist*, 26 April 2014.

19 Ori Heffetz and Katrina Ligett, "Privacy and Data-Based Research," *Journal of Economic Perspectives*, 28, no. 2 (2014): 75-98.

20 Sean Hooley and Latanya Sweeney, "Survey of Publicly Available State Health Databases," *Data Privacy Lab*, June 2013.

CHAPTER

5

ABOUT THE AUTHOR

Joel Gurin is a leading expert on Open Data—accessible public data that can drive new company development, business strategies, scientific innovation, and ventures for the public good. He is the author of the book *Open Data Now* and senior advisor at the Governance Lab at New York University, where he directs the Open Data 500 project. He previously served as chair of the White House Task Force on Smart Disclosure, as chief of the Consumer and Governmental Affairs Bureau of the U.S. Federal Communications Commission, and as editorial director and executive vice president of *Consumer Reports.*

DRIVING INNOVATION WITH OPEN DATA

BY JOEL GURIN

Key Takeaways

Open Data, like Big Data, is a major driver for innovation. Unlike privately held Big Data, Open Data can be used by anyone as a free public resource and can be used to start new businesses, gain business intelligence, and improve business processes.

While Open Data can come from many sources—including social media, private sector companies, and scientific research—the most extensive, widely used Open Data comes from government agencies and offices. Governments at all levels need to develop policies and processes to release relevant, accessible, and useful Open Data sources to enable innovation, foster a better-informed public, and create economic opportunity.

Hundreds of new companies have launched using Open Data, operating in all sectors of the economy and using a wide range of business models and revenue sources.

The chapters in this report provide ample evidence of the power of data and its business potential. But like any business resource, data is only valuable if the benefit of using it outweighs its cost. Data collection, management, distribution, quality control, and application all come at a price—a potential obstacle for companies of any size, though especially for small- and medium-sized enterprises.

Over the last several years, however, the "I" of data's return on investment (ROI) has become less of a hurdle, and new data-driven companies are developing rapidly as a result. One major reason is that governments at the federal, state, and local level are making more data available at little or no charge for the private sector and the public to use. Governments collect data of all kinds—including scientific, demographic, and financial data—at taxpayer expense.

Now, public sector agencies and departments are increasingly repaying that public investment by making their data available to all for free or at a low cost. This is Open Data. While there are still costs in putting the data to use, the growing availability of this national resource is becoming a significant driver for hundreds of new businesses. This chapter describes the growing potential of Open Data and the data-driven innovation it supports, the types of data and applications that are most promising, and the policies that will encourage innovation going forward.

Market Opportunity in Data-Driven Innovation

Today's unprecedented ability to gather, analyze, and use large amounts of data—both Open Data and privately held data—is creating qualitatively new kinds of business opportunities.

Data analysis can increase efficiency and reduce costs through what can be called *process innovation*. Logistics companies are using data to improve efficiency. UPS, for example, is using data analytics to determine the optimal routes for its drivers. A recent analysis by The Governance Lab (GovLab) at New York University, an academically based research organization, showed how data analysis can increase efficiency in healthcare, both for the National Health Service in the UK and potentially for other healthcare systems as well.[1] And as discussed throughout this report, data generated by the Internet of Things can reveal new correlations that lead to insight

and innovation. For example, data analysis of manufacturing processes can reveal opportunities to improve efficiency.

Companies can also use data to reach customers more effectively, improve brand reputation, or design products and services for specific audiences (i.e., *marketing innovation*). For example, Google analyzes search behavior to target advertising, and Amazon uses customer data to increase sales through personalized recommendations.

Lastly, data drives *business innovation and creation*. There are a growing number of companies throughout myriad industries (e.g., finance, healthcare, energy, education, etc.) that simply would not exist without today's data science. They are not using data to run their delivery network more efficiently or sell more books. They are using data to deliver entirely new products and services—to build businesses that are innovative from the ground up.

This chapter looks at this third kind of innovation, not because it will necessarily have the largest short-term economic impact but because of the impact it will have over time. These new data-driven companies will have a multiplier effect: once the first companies show the way, others will follow, leading to cumulative job growth and wealth creation. These companies rely on data as a business resource, and public government data is an especially cost-effective source for them. By making more data freely available, government agencies can make a critical difference in fostering this kind of business innovation.

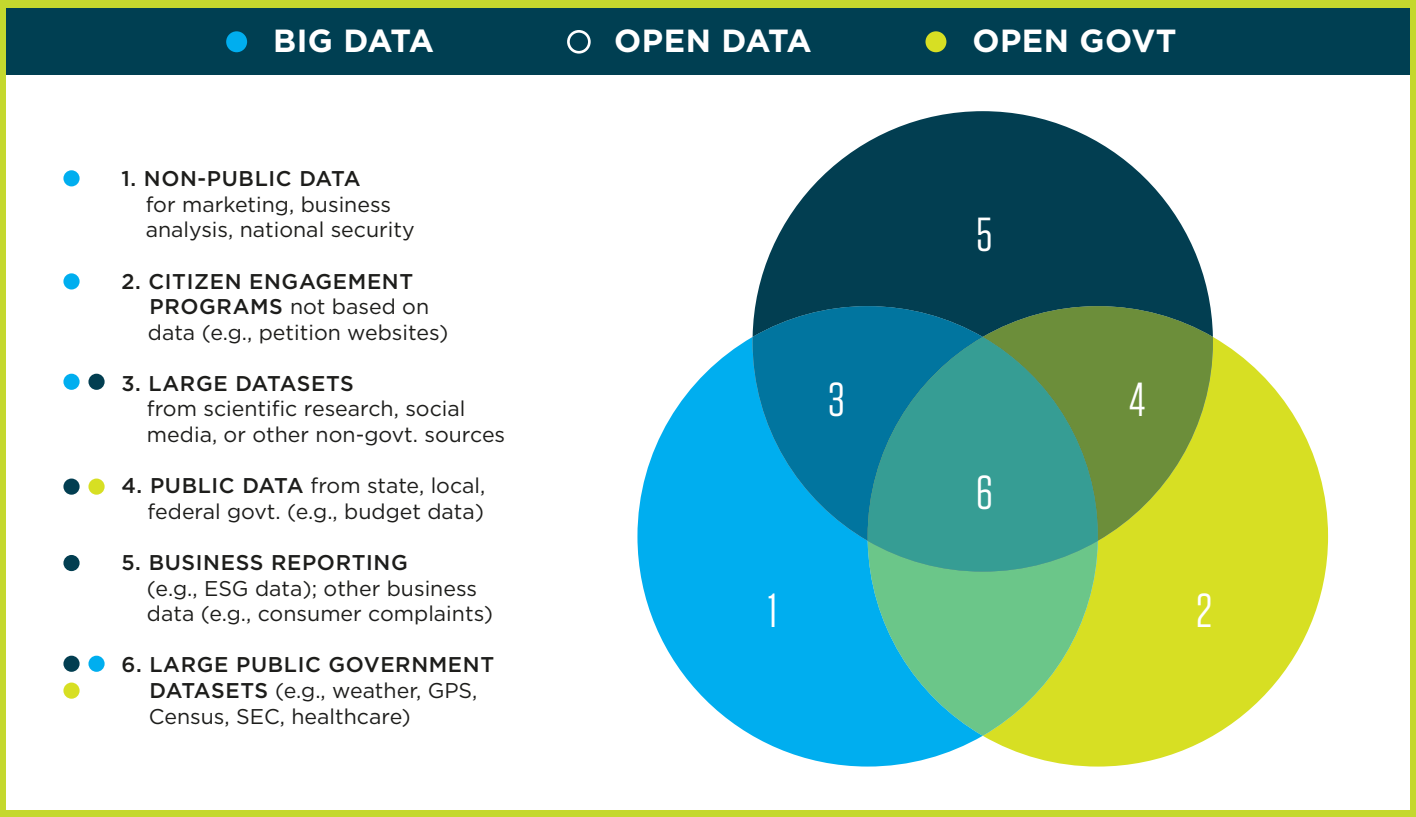
While Big Data has attracted a lot of interest, Open Data may be more important for new business creation. As the diagram below shows, Big Data and Open Data are related but different concepts. Some Big Data is anything but open. Customer records held by businesses, for example, are meant to be used exclusively by the companies that collect it to improve their business processes and marketing. Open Data, in contrast, is designed for public use. It is a public good that supports and accelerates businesses across the economy, not just specific companies in specific sectors.

When Big Data is also Open Data, as is the case for much open government data, it is especially powerful. A recent McKinsey study estimated the value of Open Data globally at more than $3 trillion a year.[2] While that study covered several kinds of Open Data, government data and large government datasets make up a significant part of that calculation. National governments around the world, with the United States and the UK in the lead, are realizing that the data they collect in areas as diverse as agriculture, finance, and population dynamics can have tremendous business value. They are now working to make those datasets more widely available, more usable, and more relevant to business needs.

Beyond open government data, three other kinds of Open Data are driving innovation in important ways:

Scientific Data
The results of scientific research have often been closely guarded. Academic researchers hold on to their data until they can publish it, while private sector research is generally not shared until the company that supported it can patent the results. Now, however, scientists in academia and business alike are beginning to test a new model, one where they share data early on to accelerate the pace of everyone's research. This open-science approach was developed most notably in the Human Genome Project, funded by the U.S. government, which ran from 1990 to 2003. The scientists involved agreed to share data openly, and that approach accelerated their progress. Now, pharmaceutical companies are beginning to experiment with a similar model of data-sharing at an early research stage.

Social Media Data
Social media is a rich source of Open Data. Between review sites, blogs, and an average of 200 billion tweets sent each year,[3] social media users are creating a huge resource of public data reflecting their opinions about consumer products, services, and brands. The evolving science of sentiment analysis uses text analytics and other approaches to synthesize those public data points into information that can be used for marketing, product development, and brand management. Companies like Gnip and Datasift have built their business on aggregating social media data and making it easy for other companies to study and analyze.



Personal Data

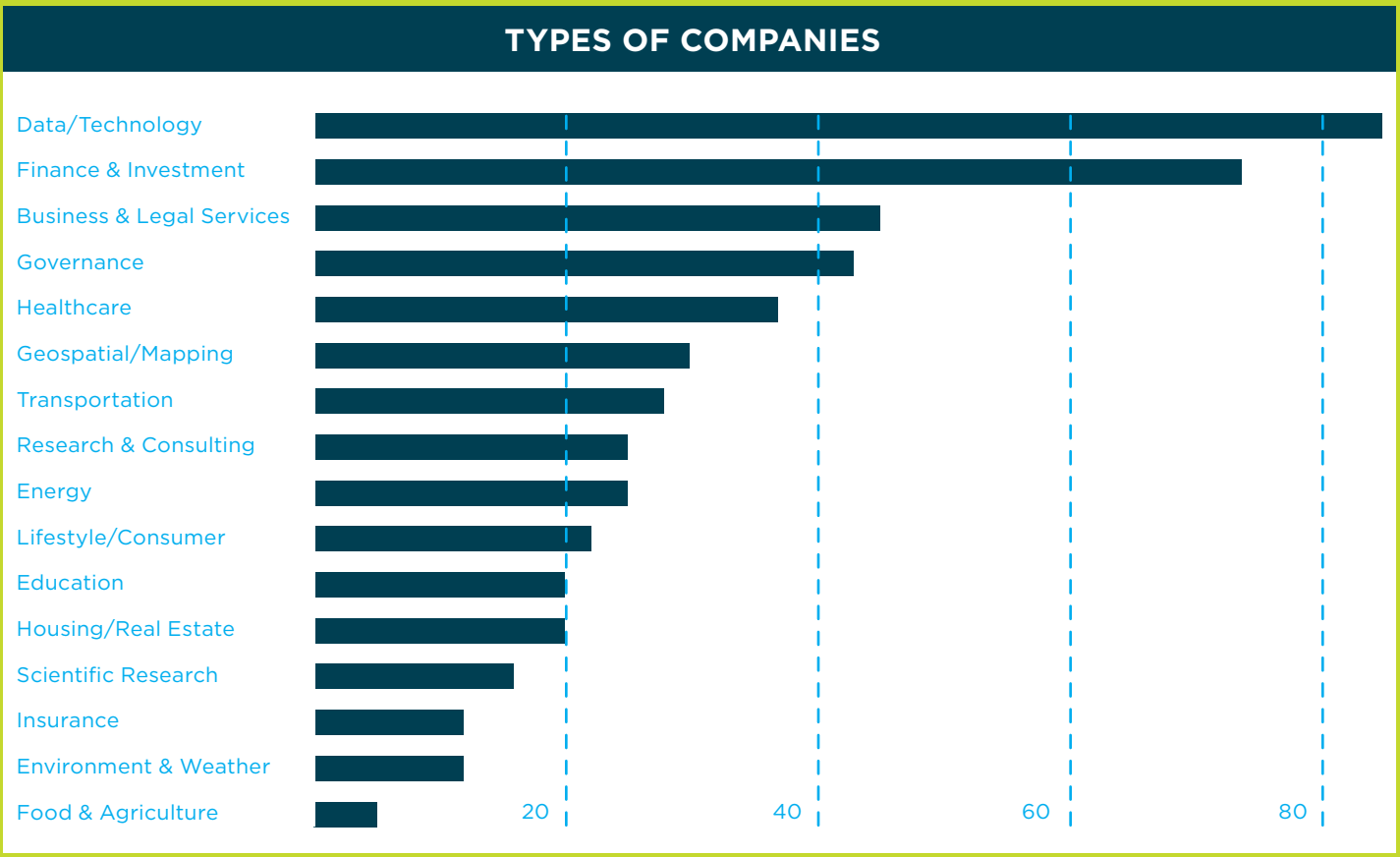
No one is suggesting that personal data on health, finances, or other individual data should be publicly available. There is increasing interest, however, in making each person's individual data more available and open to him or her. New applications are helping people download their health records, tax forms, energy usage history, and more. The model is the Blue Button program that was originally developed to help veterans download their medical histories from the Veterans Administration. The private sector has now adapted it to provide medical records to about 150 million Americans.[4] A similar program for personal energy usage data, the Green Button program, was developed through government collaboration with utilities.

Market Development — Opportunities for Using Public Data

The United States and other national governments have committed themselves to making government data "open by default;" that is, to make it open to the public unless there are security, privacy, or other compelling reasons not to do so. But datasets won't open themselves, and it is not possible to make a country's entire supply of public data available overnight. Since it will take considerable time, money, and work to turn national government datasets into usable Open Data, it is important to try to evaluate the ROI for this effort. Over the last few years, policy analysts have made several high-level attempts to estimate the economic value of these different kinds of data, Open Data in particular.

The aforementioned GovLab is studying the same issue in a more granular way. The GovLab now runs the Open Data 500 study, a project to find and study roughly 500 U.S.-based companies that use open government data as a key business resource.[5] While the study has not yet collected systematic financial data on these companies, it has provided a basic map of the territory, showing the categories of companies that use open government data, which federal agencies they draw on as data suppliers, basic information about their business models, and what kinds of open government data have the greatest potential for use.

The Open Data 500 includes companies across business sectors. Several companies are built on two classic examples of open government data: weather data, first released in the 1970s, which has fueled companies like the Weather Channel; and GPS data, made available more recently, which is used by companies ranging from OnStar to Uber. But a look at companies started in the last



10 years shows diverse uses of data from a wide range of government agencies. The table above offers about 100 examples, organized by business category. These are not meant to be a "best of" list, but rather, examples that show the types of applications in different sectors that are beginning to attract public attention and investor interest.

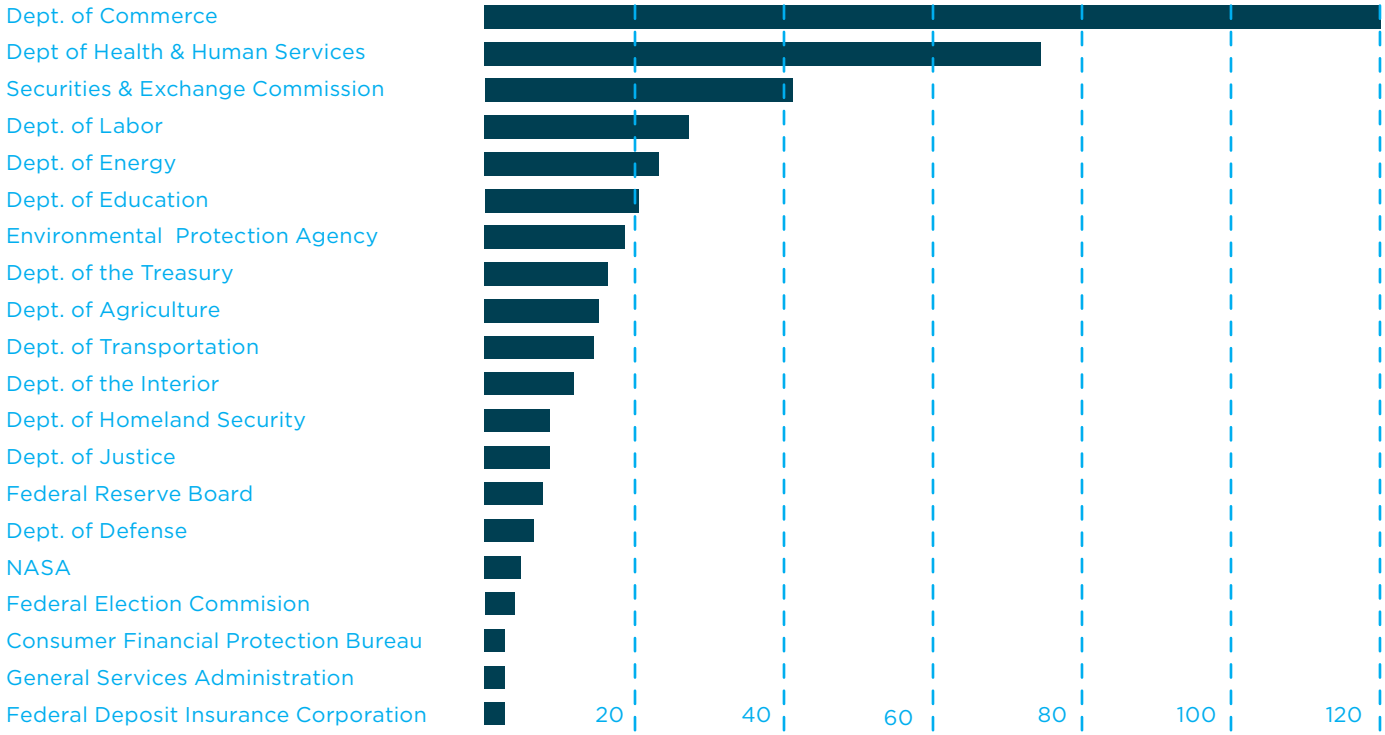
One striking development is the growing number of companies whose business is to make it easier for other businesses to use Open Data. Categorized as Data/Technology companies in the Open Data 500, they make up the largest single category in that study. These companies provide platforms and services that make open government data easier to find, understand, and use. One of the best examples is Enigma.io, a Manhattan-based company that gained visibility when it won the New York TechCrunch Disrupt competition in May 2013.[6] Enigma provides a solution for the technical limitations of government datasets by putting their data onto a common, usable platform.

These companies serve a critical function in the Open Data ecosystem. Much government data is incomplete or inaccurate, managed through obsolescent legacy systems, or difficult to find. While many government agencies are working to improve their data resources, it is a massive task and one that requires help from the private sector. Given the complexities of government datasets, the current state of much government data, and the lack of funding to improve it rapidly, companies that serve as data intermediaries will continue to have a viable business for years to come. They will also have a multiplier effect: their success will help make many other data-driven companies successful as well.

There are data-driven opportunities for businesses across all industries, with different kinds of Open Data serving as fuel for their innovative fires. Some of these most active sectors and the most important datasets include:

Business and Legal Services
A number of companies are managing, analyzing, and providing Open Data for business intelligence and business operations. Innography, for example, takes data from the U.S. Patent and Trademark Office and combines it with other data to provide analytic tools that businesses can use to learn about potential competitors and partners. In another area, Panjiva uses customs data to facilitate international trade, connecting buyers and suppliers across 190 countries.

NUMBER OF COMPANIES USING EACH AGENCIES' DATA



Education
Data-driven companies are finding value in two kinds of education data. The first is data on student performance, which can be opened to students, parents, and teachers to help tailor education to specific student needs. It is not yet clear how companies will be able to use this sensitive data to connect students with educational resources and programs without running afoul of privacy concerns. If they can, however, they will provide an important public benefit with significant economic value.

The second kind of education data is about the academic institutions themselves, and in particular, the value they offer. College-bound students have long-relied on college rankings from the likes of *U.S. News & World Report* and the *Princeton Review* to find a good college. They have used that information to work with their parents and their high school counselors to figure out whether and how they can afford the college of their choice. What many don't realize, however, is that signing up to go to any given college is like buying a new car—the sticker price is less meaningful than the price you can negotiate. Most colleges are now required to disclose their "true cost;" that is, the expected cost for a particular kind of student after the college's typical financial aid package is taken into account. Several education websites are now using this kind of data to help students find colleges and perhaps find a school that is more affordable than they thought.

Energy
Growing interest in clean energy and sustainability is creating a new breed of data-driven companies. Several now use a combination of Open Data on energy efficiency with data-gathering sensors to help make residential and commercial buildings more energy efficient. Others are using Open Data to help advance clean energy technologies. Clean Power Finance, for example, uses Open Data to help solar-power professionals find access to financing. A new company, Solar Census, is aiming to make solar power more cost-effective by using geospatial and other data to figure out exactly how to place and position solar panels for maximum efficiency.

Finance and Investment
This is perhaps the most developed category of Open Data businesses, as finance and investment companies have long-used open government data as an essential resource. Data from the Securities and Exchange Commission (SEC) has

powered investment firms for decades, and it is now possible to combine SEC data with other data sources for faster, more accurate, and more usable analysis. For example, Analytix Insight, which runs the website Capital Cube, provides analyses of more than 40,000 publicly traded global companies, updated daily, and presented in formats that make it easy for investors to use.

Other new companies provide a wide range of financial information and services to businesses and consumers. Brightscope uses information filed with the Department of Labor to evaluate the fees charged by different pension plans and helps employers and employees make more informed choices. Companies like Credit Sesame and NerdWallet compare different options and recommend credit cards and other financial services to consumers based on their credit ratings. Bill Guard uses data from the Consumer Financial Protection Bureau and information submitted by consumers to help protect people from fraudulent charges.

Some financial information companies are now processing financial data in the interest of helping small- and medium-sized enterprises (SMEs) get the capital they need—another example of the multiplier effect. These companies have realized that SMEs suffer because lenders cannot afford to do due diligence for small companies and thus don't have the confidence to give them the funds they need. On Deck now uses a number of public data sources to do that risk assessment and help small businesses get access to much-needed business loans. In a similar way, the British company Duedil serves to facilitate funding for SMEs in the UK and Ireland.

Food and Agriculture

In this area, perhaps more than any other sector besides healthcare, Open Data has the potential to revolutionize an industry that is essential to society and human wellbeing. The Climate Corporation, an iconic example of a successful Open Data company, has pioneered what is now being called "precision agriculture"—using Open Data to help farmers increase their efficiency and the profitability of their farms. The Climate Corporation, which was sold to Monsanto in the fall of 2013 for about $1 billion, built value by combining different Open Data sources (ranging from satellite data to information on rainfall and soil quality) and subjecting it to sophisticated analysis.[7] The result is a set of services that can help farmers decide which crops to plant and when and help them prepare for the impact of climate change. Other companies, like FarmLogs, are beginning to offer some similar services.

Governance

Local government data is often no easier to use than federal data. Different cities use different and often unwieldy systems to track their government operations. Companies like OpenGov and Govini are providing platforms that municipal governments can use to organize their data and share it with their citizens. Organized and presented in clear charts and graphs, city data can become a tool for town meetings, city planning, and dialogue with city leaders. These tools also make it possible to compare operations in similar cities. For example, local data can allow a comparison between police overtime hours in Palo Alto and those in San Mateo, potentially revealing the reason for any disparity.

Housing and Real Estate

Real estate websites (which emerged about a decade ago) do much more than aggregate listings from brokers. Sites like Redfin, Trulia, and Zillow now offer data on schools, walkability scores, crime rates, and many other quality of life indicators, using data from national and local sources. In a country where historical averages show about one-fifth of the population moving every year, we can expect these sites to compete increasingly on the depth of information they offer and their ease of use.

Lifestyle and Consumer

In May 2013, the White House released the report of the Task Force on Smart Disclosure, a group chaired by this author to study how open government data can be used for consumer decision making.[8] Federal agencies like the Consumer Financial Protection Bureau, the Department of Health and Human Services, and others now have data on a wide range of consumer services, including credit cards, mortgages, healthcare services, and more. Websites like FindTheBest have begun to use this kind of data to provide consumer guidance on a range of products and services.

As the idea of Smart Disclosure takes hold, we can expect to see more websites tailored to particular consumer needs and concerns. GoodGuide, for

example, uses data from more than 1,500 datasets to create a service for consumers who want to choose the products they buy with an eye towards their environmental impact, health concerns, or other factors. GoodGuide's analysis is not only being used by consumers but also by companies that want to use the data to "go green."

Transportation
The availability of new, usable transportation data is transforming this sector as well. The applications include companies that provide detailed directions and traffic advisories (HopStop, Roadify Transit), traffic analytics to help transportation planners (Inrix Traffic), and safety data to improve the trucking industry (Keychain Logistics).

The Biggest Opportunity for Data-Driven Disruption — Healthcare

While all these categories of data-driven companies have significant growth potential, it is in healthcare that new uses of data may bring the greatest opportunity for disruptive innovation. We can expect more efficient systems for tracking patients and their care, leading to lower costs and fewer medical errors. We can look forward to more data-driven diagnostics, treatment plans, and predictive analytics to more scientifically determine the best treatments. And we will see a new era of personalized medicine, where data about an individual—ranging from genetic makeup to exercise habits—is used to algorithmically determine a strategy for care.

Healthcare has become a proving ground that shows how the four different kinds of data—Big Data, Open Data, personal data, and scientific data—can be used together to great effect. By analyzing *Big Data* (the voluminous information on public health, treatment outcomes, and individual patient records), healthcare analysts are now able to find patterns in public health, healthcare costs, regional differences in care, and more. *Open Data* on healthcare is becoming more available through the Centers for Medicare and Medicaid Services (CMS) and recent data releases from the U.S. Food and Drug Administration. With *personal data*, the third piece of the puzzle, people are getting more data to help them understand and manage their own health issues, both through Blue Button and similar programs and through personal health monitoring devices. And we're seeing a rapid increase in open *scientific data*, particularly data about the human genome, which can be used to improve medical care.

New companies are launching to put all this data to work. Venture capitalists reportedly invested more than $2 billion in digital health startups in the first half of 2014.[9] There are three categories of companies that are growing most rapidly.

Healthcare Selection
A number of websites now use a combination of Open Data and consumer feedback to provide information on the quality and cost of different healthcare options. ZocDoc and Vitals help people find doctors and clinics and book appointments. Aidin uses data from CMS and other sources to help hospital discharge planners work with their patients to find better post-hospital care. Drawing on Open Data from the U.S. National Provider Identifier Registry, iTriage lets you use a website or smartphone to log in symptoms, get quick advice on the kind of care needed, and get a list of nearby facilities that can help. And TrialX connects patients with clinical trials of new treatments. As CMS releases more data on both the quality and cost of care, data-driven healthcare companies will have an opportunity to help individuals and drive down national healthcare costs.

Personal Health Management
The movement to electronic medical records will open new ways for individuals and their doctors to combine public and personal information to improve their healthcare. Amida Technology Solutions is building on the Blue Button model to accelerate the use of personal health records. At the same time, other companies are tapping the power of personal data in different ways. Propeller Health uses inhaler sensors, mobile apps, and data analytics to help doctors identify asthma patients who need additional help to control their chronic disease. Iodine combines large healthcare datasets with individualized health information to provide patient guidance. And several companies have developed wristbands and other wearable monitors that track personal biometric data as an aid to wellness programs and medical treatment.

Data Management and Analytics
As more health data is opened up, more companies are finding ways to analyze it. Evidera uses data from CMS, databases of clinical trials, and other sources to develop models predicting how different treatment interventions will affect different kinds of patients. In a similar way, Predilytics uses machine learning to help health plans and providers deliver care more effectively

and reduce costly admissions (and readmissions) to the hospital.

Business and Revenue Models for Data-Driven Companies

Open Data poses a business paradox. How can one hope to build a business worth millions or even billions of dollars by using data that is free to the public? Open Data startups have succeeded by bringing new ideas, analytic capabilities, user-focused design, and other added value to the basic value inherent in Open Data. As with many startups, the revenue model for many of these companies is still a work in progress. They have focused on functionality first, monetization second. (In a few years, we will know whether this has been a wise strategy.) Nevertheless, several business models are starting to emerge.

In a 2012 study,[10] Deloitte surveyed a large sample of Open Data companies and identified five business archetypes:

Suppliers publish Open Data that can be easily used.

Aggregators collect Open Data, analyze it, and charge for their insights or make money from the data in other ways.

Developers "design, build, and sell Web-based, tablet, or smart-phone applications" using Open Data as a free resource.

Enrichers are "typically large, established businesses" that use Open Data to "enhance their existing products and services," for example, by using demographic data to better understand their customers.

Enablers charge companies to make it easier for them to use Open Data.

The Open Data 500 study has found a number of companies that combine several Deloitte archetypes, particularly among companies that the Open Data 500 categorizes as "Data/Technology." For example, Enigma.io, as described above, has *aggregated* about 100,000 government datasets, *supplied* that data to the public in a more useful form, and served as an *enabler* by consulting with companies that have special uses for certain kinds of datasets (e.g., risk analysis).

While Deloitte's categories describe the different ways in which companies use Open Data to deliver business value, the Open Data 500 study has focused on a different part of the business model—the ways companies generate revenue from their work. The Open Data 500 has found a variety of revenue sources that are available to companies across the archetypes identified by Deloitte.

Advertising, a common revenue source for websites, may be a good source for some Open Data companies. Websites are increasingly relying on "native advertising;" that is, sponsored content that is written at an advertiser's direction but presented in a way that looks like the site's regular content. This kind of advertising, however, may be at odds with Open Data companies that base their business on the promise of providing objective and unbiased information.

Subscription models, in contrast, can be a natural fit for these data-driven companies. Many add value to Open Data as they combine datasets, analyze data, visualize it, or present it in ways that are tailored to the user's needs. Willingness to pay depends on the relevance, complexity, uniqueness, and value of the information. While a consumer would be unlikely to subscribe to a simple website that helps him or her choose a credit card, a farmer could easily find it worthwhile to subscribe to a service that uses data to help improve the farm's profitability.

Lead generation is another natural revenue source for data-driven companies that evaluate business services or help consumers find products and services. Real estate sites are a classic example. They collect a broker's fee when someone uses the website to find and buy a home. The challenge with this revenue model, however, is that it can give companies an incentive to game the system and refer consumers to service providers that pay the highest referral fees. Over time, this model could generate consumer distrust and become less effective. Companies that use this revenue source should consider establishing a voluntary code of conduct that would include transparency about their business models—enough to let users know how the company earns its revenue and give them the assurance that their information is unbiased.

Fees for data management and analytics provide revenue to companies that help clients learn more from the data available to them. They may work

REVENUE SOURCES FOR OPEN DATA COMPANIES

with businesses, governments, or both. Several companies now help government agencies manage and analyze their own data—or even sell agencies' data back to them in an improved form, as Panjiva does with customs data from the federal government.

Consulting fees are yet another revenue source. Some data-driven firms, like Booz Allen and McKinsey, analyze both open and proprietary data to advise their corporate clients on business opportunities, while investment firms use increasingly diverse data sources to predict market trends.

Finally, licensing fees are a source of revenue for the kinds of companies that Deloitte calls enablers. They can license software, tools, platforms, database services, cloud-based services, and more to enable new data-driven companies to build their business.

Potential Barriers and Ways to Overcome Them

Any discussion of data-driven business has to deal with the issue of data privacy. Chapter 7 addresses privacy concerns surrounding the use of proprietary Big Data, such as personal data gathered online, through data brokers, through customer records, or in other ways. Companies driven by Open Data do not generally use this kind of personal information, but they may need to access datasets that aggregate personal data and present it in a way that masks personal information. Healthcare companies, for example, may want to use anonymous patient records in a way that enables them to detect patterns in treatment outcomes, such as correlations between prescription drugs, lifestyle, and therapeutic results.

There is an ongoing debate about whether it is truly possible to anonymize data like this or whether any system of anonymization can

ultimately be defeated. This debate is likely to play out over the next few years and could have a major impact on data-driven innovation. If successful technologies for anonymization are developed, they will open up new opportunities for data analysis and publication. On the other hand, if experts and the public come to believe that individuals' identities can always be deciphered from the data, then a number of paths to innovation will be cut off.

Data-driven businesses also have to deal with one of the biggest obstacles to growth: poor-quality data. The quality of U.S. government data varies greatly between agencies and sometimes even within the same agency. Government data systems have grown by accretion over decades. They are often housed in obsolete data management systems and, at the same time, may have errors, gaps in the data, or out-of-date information. These are not easy problems for either the government or third parties to solve, but without some solutions, the potential of open government data will be underused.

Government agencies that provide data, and the businesses and non-profits that use it, all have a common interest in making government data as relevant, accessible, actionable, and accurate as possible. None can do it acting alone. What is needed is a way to bring together data providers with data users for a structured, action-oriented dialogue to identify the most important datasets for business and public use and find ways to improve them.

The GovLab has launched a series of Open Data Roundtables to bring together federal agencies with the businesses that use their data. The first such roundtable, held with the Department of Commerce in June 2014, included more than 20 officials and staff from the department and about 20 businesses. This event was the beginning of a process designed to identify specific areas for improving data quality and accessibility. It was followed by an Open Data Roundtable with the U.S. Department of Agriculture. As of this writing, additional GovLab roundtables are being planned with the Departments of Labor, Transportation, and Treasury, as well as with other federal agencies.

A next step could be to develop public-private collaborations to turn government data into machine-readable forms that could make it much more useful and help drive innovation. Some companies, such as Captricity, have developed the technology to convert data from PDF files (a document type common in government data) into more usable formats. By working together, government agencies and these companies could convert large amounts of the most important data into a form that other companies could easily use.

Supporting data-driven innovation will also require government policies that make new sources of data available and encourage companies to use them. The federal Open Data Policy,[11] established in May 2013, and the National Open Data Action Plan,[12] released one year later, set out some basic principles for the U.S. government to use in promoting Open Data. The Open Data Policy not only directs federal agencies to release more Open Data, it also requires them to release information about data quality. We can hope and expect that they will do some data cleanup themselves, demand better data from the businesses they regulate, or use creative solutions, like turning to crowdsourcing for help.

The federal government has steadily made Data.gov (the central repository of its Open Data) more accessible and useful. The General Services Administration, which administers Data.gov, plans to keep working to make this key website better still. As part of implementing the Open Data Policy, the administration has also set up Project Open Data on GitHub, the world's largest community for open-source software. These resources will be helpful for anyone working with Open Data, be it inside or outside of government.

As more and better Open Data becomes available, we will learn more about the best ways to use it for business applications, job creation, and economic growth. While it is clear that Open Data can be used in a wide range of industries, we do not yet know exactly which kinds of applications will turn out to be the most promising, the most robust, and the most replicable. We need to learn more about the mechanisms of value creation using Open Data, the kinds of Open Data that will be most important to various sectors, and the ways in which Open

Data fits into different companies' strategic and operating models. Ongoing research on the uses of Open Data by academic institutions, government agencies, and independent organizations will be essential to ensure that our public data resources are used widely and well.

Ultimately, creating a better Open Data ecosystem will take both public and private resources and funding. The payback for government technology improvements is generally calculated based on short-term savings, like improvements in efficiency and cost reduction. But the release of more and better open government data can have economic benefits that multiply over time. An investment in Open Data now will pay off for years to come.

ENDNOTES

1 Stefaan Verhulst, "The Open Data Era in Health and Social Care," *The Governance Lab*, May 2014.

2 James Manyika et al., "Open Data: Unlocking Innovation and Performance with Liquid Information," *McKinsey Global Institute*, Oct. 2013.

3 Internet Live Stats, "Twitter Usage Statistics" <http://www.internetlivestats.com/twitter-statistics/> (19 Aug. 2014).

4 Nick Sinai and Adam Dole, "Leading Pharmacies and Retailers Join Blue Button Initiative," *Health IT Buzz*, 7 Feb. 2014.

5 The author of this chapter is a senior advisor at The GovLab and project director for Open Data 500.

6 Megan Rose Dickey, "This Tech Crunch Disrupt Winner Could Be the Future of Search," *Business Insider*, 2 May 2013.

7 "Monsanto Completes Acquisition of The Climate Corporation," *MarketWatch*, 1 Nov. 2013.

8 Executive Office of the President, National Science and Technology Council, "Smart Disclosure and Consumer Decision Making: Report of the Task Force on Smart Disclosure," May 2013.

9 Eric Whitney, "Power to the Health Data Geeks," *National Public Radio*, 16 June 2014.

10 "Open Growth: Stimulating Demand for Open Data in the UK," *Deloitte Analytics*, 2012.

11 OMB Memorandum for the Heads of Executive Departments and Agencies M-13-13, "Open Data Policy – Managing Information as an Asset," 9 May 2013.

12 The White House, "U.S. Open Data Action Plan," 9 May 2014.

CHAPTER
6

ABOUT THE AUTHORS

Benjamin Wittes is a Senior Fellow in Governance Studies at The Brookings Institution. He co-founded and is the editor-in-chief of the Lawfare blog, which is devoted to sober and serious discussion of "hard national security choices," and is a member of the Hoover Institution's Task Force on National Security and Law. He is the author of *Detention and Denial: The Case for Candor After Guantanamo,* published in November 2011, co-editor of *Constitution 3.0: Freedom and Technological Change,* published in December 2011, and editor of *Campaign 2012: Twelve Independent Ideas for Improving American Public Policy* (Brookings Institution Press, May 2012). He is also writing a book on data and technology proliferation and their implications for security. He is the author of *Law and the Long War: The Future of Justice in the Age of Terror,* published in June 2008 by The Penguin Press, and the editor of the 2009 Brookings book, *Legislating the War on Terror: An Agenda for Reform.*

Wells C. Bennett is a Fellow in the Brookings Institution's Governance Studies program, and Managing Editor of Lawfare, a leading web resource for rigorous, non-ideological analysis of "Hard National Security Choices." He concentrates on issues at the intersection of law and national security, including the detention and trial of suspected terrorists, targeted killing, privacy, domestic drones, Big Data, and surveillance.

DATABUSE IN THE BIG DATA ERA

BY BENJAMIN WITTES & WELLS C. BENNETT

Key Takeaways

Privacy is actually not a great vocabulary for discussing corporate responsibilities and consumer protection. The word promises a great deal more than policymakers are prepared to deliver, and in some ways, it also promises more than consumers want.

Protection against databuse—the malicious, reckless, negligent, or unjustified handling, collection, or use of a person's data in a fashion adverse to that person's interests and in the absence of that person's knowing consent—should lie at the core of the relationship between individuals and the companies to whom they give data in exchange for services.

Companies must be reasonable and honest custodians—trustees—of the material we have put in their hands. This means handling data in an honest, secure, and straightforward fashion, one that does not injure consumers and that gives them reasonable information about and control over what is and is not being done with the data they provide.

The following paper is an abridged version of "Databuse and a Trusteeship Model of Consumer Protection in the Big Data Era," published by The Brookings Institution on June 4, 2014.

How much does the relationship between individuals and the companies in which they entrust their data depend on the concept of "privacy?" And how much does the idea of privacy really tell us about what the government does, or ought to do, in seeking to shield consumers from Big Data harms?

There is reason to ask. Privacy is undeniably a deep value in our liberal society. But one can acknowledge its significance and its durability while also acknowledging its malleability. For privacy is also something of an intellectual rabbit hole, a notion so contested and ill-defined that it often offers little guidance to policymakers concerning the uses of personal information they should encourage, discourage, or forbid. Debates over privacy often descend into an angels-on-the-head-of-a-pin discussion. Groups organize around privacy. Companies speak reverently of privacy and have elaborate policies to deliver it—or to justify their handling of consumer data as consistent with it. Government officials commit to protecting privacy, even in the course of conducting massive surveillance programs. And we have come to expect as much, given the disagreement in many quarters over what privacy means. The invocation of privacy mostly serves to shift discussion, from announcing a value to addressing what that value requires. Privacy can tell a government or company what to name a certain policy after, but it doesn't answer many questions about how that company or government ought to handle that data.

Moreover, in its broadest conception, privacy also has a way of overpromising—of creating consumer expectations on which our market and political system will not, in fact, deliver. The term covers such a huge range of ground that it can, at times, suggest protections in excess of what regulators are empowered to enforce by law, what legislators are proposing, and what companies are willing to provide consistent with their business models.

In 2011, in a paper entitled "Databuse: Digital Privacy and the Mosaic," one of us suggested that "technology's advance and the proliferation of personal data in the hands of third parties has left us with a conceptually outmoded debate, whose reliance on the concept of privacy does not usefully guide the public policy questions we face."

Instead, the paper proposed thinking about massive individual data held in the hands of third-party companies with reference to a concept it termed "databuse," which it defined as: "the malicious, reckless, negligent, or unjustified handling, collection, or use of a person's data in a fashion adverse to that person's interests and in the absence of that person's knowing consent."

Databuse, the paper argued, "can occur in corporate, government, or individual handling of data. Our expectations against it are an assertion of a negative right, not a positive one. It is in some respects closer to the non-self-incrimination value of the Fifth Amendment than to the privacy value of the Fourth Amendment. It asks not to be left alone, only that we not be forced to be the agents of our own injury when we entrust our data to others."[1]

We attempt in this essay to sketch out the data protection obligations that businesses owe to their users. We attempt to identify, amid the enormous range of values and proposed protections that people often stuff into privacy's capacious shell, a core of user protections that actually represent something like a consensus.

The values and duties that make up this consensus describe a relationship best seen as a form of trusteeship. A user's entrusting his or her personal data to a company in exchange for a service, we shall argue, conveys certain obligations to the corporate custodians of that person's data: obligations to keep it secure; obligations to be candid and straightforward with users about how their data is being used; obligations not to materially misrepresent their uses of user data; and obligations not to use them in fashions injurious to or materially adverse to the users' interests without their explicit consent. These obligations show up in nearly all privacy codes, in patterns of government enforcement, and in the privacy policies of the largest Internet companies. It is failures of this sort of data trusteeship that we define as databuse. And we argue that protection against databuse—and not broader protections of more expansive, aspirational visions of privacy—should lie at the core of the relationship between individuals and the companies to whom they give data in exchange for services.

Privacy, Trusteeship, And Databuse

Our premise is straightforward: "privacy," while a pervasive rhetoric in the area of data handling and management, is actually not a great vocabulary for discussing corporate responsibilities and consumer protection. The word promises a great deal more than policymakers are prepared to deliver, and in some ways, it also promises more than consumers want.

The concept certainly was not inevitable as the reference point for discussions of individual rights in the handling of data. It developed over time, in response to the obsolescence of previous legal constructions designed to shield individuals from government and one another. To put the matter simply, we created privacy because technology left previous doctrines unable to describe the intrusions on our seclusion that we were feeling.

Ironically, today it is privacy itself that no longer adequately describes the violations people experience with respect to large caches of personal data held by others—and it describes those violations less and less well as time goes on. Much of the material that makes up these datasets, after all, involves records of events that take place in public, not in private. Much of this data is sensitive only in aggregation; it is often trivial in and of itself—and we consequently think little of giving it, or the rights to use it, away.

When one stops and contemplates what genuinely upsets us in the marketplace, broad conceptions of privacy—conceptions based on secrecy or non-disclosure of one's data—do not express it well at all. It's not just that we happily trade confidentiality and anonymity for convenience. It's that we seem to have no trouble with disclosures and uses of our data when they take place for our benefit. We do not punish companies that aggressively use our data for purposes of their own, so long as those uses do not cause us adverse consequences.

Were we truly concerned with the idea that another person has knowledge of these transactions, we would react to these and many other routine online actions with more hostility.

"THE IDEA OF TRUSTEESHIP IS CENTRAL HERE, IN THAT IT HELPS GUIDE BOTH CONSUMER EXPECTATIONS AND CORPORATE BEHAVIOR."

Yet, we have no trouble with outside entities handling, managing, and even using our data—as long as we derive some benefit or, at least, incur no harm as a result. Rather, we positively expect uses of our data that will benefit or protect us; we tolerate uses of them so long as the consequences to us are benign; and we object viscerally only where the use of our data has some adverse consequence for us. This is not traditional privacy we are asking for. It is something different. That something is protection against what we call databuse.

Think of databuse as that core of the privacy spectrum that is most modest in nature. Databuse is different from broader visions of privacy in that it does not presume as a starting point the non-disclosure, non-use, even quarantining from human eyes of data we have willingly transacted in exchange for services.[2] It instead treats the dissemination of such data—in whole or in part—as an option we might or might not want to choose.

Databuse asks only for protection against unwarranted harms associated with entrusting our data to large entities in exchange for services from them. It asks that the costs of our engagement with these companies not be a total loss of control of the bits and pieces of data that make up the fabric of our day-to-day lives. It asks, in short, that the companies be reasonable and honest custodians—trustees—of the material we have put in their hands. It acknowledges that they will use it for their own purposes. It asks only that those purposes do not conflict with our own purposes or come at our expense.

The idea of trusteeship is central here, in that it helps guide both consumer expectations and corporate behavior. A trustee in the usual sense is supposed to be a good steward of property belonging to somebody else. That means an obligation, first and foremost, to administer the trust in the interest of the beneficiary, according to the trust instrument's terms.[3] A trustee is bound to act prudently, with reasonable care, skill, and caution,[4] and to keep beneficiaries reasonably informed, both about the trust's formation and its subsequent activities—including any changes to the trust's operation.[5]

The analogy between trusts and data-driven companies is, of course, imprecise. Facebook—as custodian of your data—is not under any obligation to act in your financial interests or to take only actions with your best interests in mind. You do not expect that. The essence of this sort of data trusteeship is an obligation on the part of companies to handle data in an honest, secure, and straightforward fashion, one that does not injure consumers and that gives them reasonable information about and control over what is and is not being done with the data they provide.

This can be teased out into distinct components. These components are familiar enough, given the policy world's long-running effort to convert vague privacy ideas into workable codes of behavior. That project can be traced back at least to the Fair Information Practice Principles ("FIPPs"), which were themselves largely derived from a 1973 report by the Department of Health, Education and Welfare on "Records, Computers, and the Rights of Citizens." In the years since, scores of articles, privacy policies, and government documents have drawn on the FIPPs. Recently, the Obama administration has relied upon them in ticking off a checklist of do's and don'ts for companies holding significant volumes of consumer data.[6] Our own catalog of corporate responsibilities broadly overlaps with that of the Obama administration—and the government studies and reports and academic literature the administration has relied upon. As we show, they also reflect the set of expectations that, when companies fail to meet them, yield enforcement actions by the Federal Trade Commission (FTC). And they also reflect the commitments the major data-handling companies

actually make to their users. In other words, the components of databuse are the parts of privacy about which we all basically agree.

To name but a few of the consensus-backed principles, first, companies must take responsibility for the secure storage, custody, and handling of personal data so that the consumer is actually providing data only to those entities to which he or she actually agrees to give them.[7] Data breaches are a major source of risk for consumers, the cause of identity thefts, fraud, and all kinds of scams. Protecting data is no less an obligation for a company that asks individuals to entrust it with data than it is for a bank that asks for trust in storing private money.

Second, companies must never use consumer data in a fashion prejudicial to the interests of consumers. Consumers are far savvier than some privacy advocates imagine them to be, and we believe individuals are generally capable of making reasonable risk-management choices about when to trade personal data in exchange for services of value. These risk-management decisions, however, require a certain faith that businesses in question—while pursuing interests of their own—are not actively subverting the consumers' interests.

This point is complicated because not everyone agrees about what it means to act in a fashion prejudicial to someone's interests. For that reason, it is critical to let individuals make their own choices both about whether to do business with a given company and, to the maximum extent possible, about what they do and do not permit that company to do with their data.

That means, third, requiring honest and straightforward accounts by companies of how they use consumer data: what they do with it; how they monetize it; what they do not do with it.[8] This does not mean an endless, legalistic "Terms of Service" document that nobody reads but simply clicks through. Such documents may be important from the standpoint of technical compliance with the law, but they do not reasonably inform the average consumer about what he can or cannot expect. Rather, it means simple, straightforward accounts of what use the company makes of consumer data. It also means not retroactively changing those rules and giving consumers reasonable notice when rules and defaults are going to change prospectively. Companies differ quite a bit in the degree of useful disclosure they give their users—and in the simplicity of those disclosures. Google and Facebook, for instance, have both created useful and simple disclosure pages. Other companies provide less information or obscure it more.

Fourth, it also means—to the maximum extent possible—giving consumers control over those decisions as applied to them.[9] This is not a binary rule, consumer control not being an on-off switch. It is a spectrum, and again, companies differ in the degree to which they give consumers control over the manner in which they use those consumers' data. Facebook now gives users fairly specific control over whom they want to share materials with.[10] Google offers users remarkably granular control over what sort of advertising they do and don't want to see and to what extent they want advertising based on their recorded interests.[11] The more control consumers have over who has

"PROTECTING DATA IS NO LESS AN OBLIGATION FOR A COMPANY THAT ASKS INDIVIDUALS TO ENTRUST IT WITH DATA THAN IT IS FOR A BANK THAT ASKS FOR TRUST IN STORING PRIVATE MONEY."

access to their data and what the trustee company can do with it, the less capacity for databuse the relationship with that company has.

Finally, fifth, companies have an obligation to honor the commitments they make to consumers regarding the handling of their data. Promising a whole lot of user control is worthless if the promises are not honored. And a great many of the FTC's enforcement actions naturally involve allegations of companies committing themselves to a set of practices and then failing to live up to them.[12]

Notice how much of conventional privacy this conception leaves out. For starters, it leaves out the way we feel when information about us is available to strangers and the sense that, quite apart from any tangible damage a disclosure might do us, our data is nobody else's business. "Privacy as sentiment" is central to much of the privacy literature today and has often played a role in the way the FTC talks about the subject, particularly with respect to its authority to urge best practices. It plays a huge role in European attitudes towards privacy. A related conception of privacy sees in it some kind of right against targeted advertising and behavioral profiling—at least in its more aggressive forms. And many commentators see in privacy as well some right to control our reputations.

At least as to companies with which the user has a direct relationship, the databuse conception largely throws this out. It requires honest, straightforward dealings by companies. It requires that the user have fair and reasonable opportunity to assess the impact on values and interests she might care about—privacy among them—of giving over her data to the company. But it ultimately acknowledges that targeted advertising is something she might want, or something she might not mind, and it considers her reputation ultimately her own responsibility to protect.

Interests Congruent And Conflicting

One simple way to think about the spectrum between good trusteeship and databuse is to examine the similarity or conflict between a consumer's interests and the company's interests in the handling of that consumer's data. Not all such uses are objectionable. Many are beneficial to the consumer, the very essence of the service the company provides. We do business with Facebook and Twitter, after all, so they can share our data with our friends and people who are interested in what we have to say. Google Maps can tell you what roads are congested because lots of phones are sending it geolocation data—phones that may well include yours. Some uses of our data, in other words, actively serve or support our interests. By contrast, a company that collects consumer data in the course of providing a service and then monetizes that data in a fashion that exposes consumers to risks they didn't reasonably bargain for is a wholly different animal.

So let's consider three different general categories of data use by companies with direct relationships with their customers.

Category I involves situations in which the consumer's interests and the company's interests align. A company wants to use the data for a particular purpose, and a consumer either actively wants the company to use the data for that purpose or actively wants services that depend pervasively on those uses of data.

This first grouping derives in part from consumers' motivations for offering up their data in the first place. People sign up for Google applications, for example, for many different reasons. But certainly among them are obtaining a convenient mechanism for sending and receiving electronic mail through the cloud, searching the Web, and figuring out, in real time, the fastest travel routes from one place to another while avoiding accidents or high-traffic areas. All of these services necessarily entail a certain measure of data usage and processing by the company to which the data is given: a message's metadata must be utilized and its contents electronically repackaged to facilitate the message's transmission from sender to intended recipient. And in order to carry out its mission of directing you from one place to another, Google Maps likewise must obtain and compare your location to the underlying map and to data identifying bottlenecks, roadblocks, or other trip-relevant events—data it is often getting by providing similar services to other users. Another everyday example, this one involving a common commercial exchange: most people use credit cards, either for the convenience or to borrow money from the issuing banks, or both. The bank, in turn, periodically scans customer accounts—peeking at the patterns of transactions—for activity indicative of possible theft or fraud. Most consumers actively want these services.

DATA USAGE CATEGORIES

CATEGORY I

Company and Consumer Data Usage Interests Align

CATEGORY II

Data Usage Advances Company's Interests, Neither Advances Nor Undercuts the Consumer's Interests

CATEGORY III

Data Usage Advances Company's Interests, Abuses Consumer's interests

The foregoing class of data handling manages to advance both parties' interests, and in obvious ways. Because of Google's practices, the customer gets better service from Google—or in some cases gets service at all. In critical respects, this is often not the use of data as a currency in exchange for the service. This is the use of data in order to provide the service. Similarly, in our banking hypothetical, snooping around for fraud and money-laundering reassures and protects the consumer for so long as she has entrusted her hard-earned cash—and the data about her transactions—to Bank of America, for example.

Category I data use thus results in an easily identifiable, win-win outcome for company and consumer alike. The tighter the link between a given use and the reason a user opts to fork over his data in the first place, the more likely that use is to fall within Category I. Category I generally does not raise the hackles of privacy advocates, and the pure Category I situation generally ought to draw the most minimal attention from policymakers, as well as impose only the most minimal corporate duty to apprise the consumer of the details surrounding its activity. By way of illustration, UPS need not obtain permission before performing any electronic processes necessary to ensure a package's safe delivery; and PayPal likewise doesn't have to ask before it deploys its users' data in an exercise meant to beta test its latest security protocols.[13]

There are, of course, legitimate questions about the boundaries of Category I with respect to different companies. Some companies would argue that advertising activities should fall within Category I, as they are making money by matching consumers with goods and services they want to purchase. For some consumers, particularly with respect to companies whose products they particularly like, this may even be correct. Many people find Amazon's book recommendations, based on the customer's prior purchasing patterns, useful, after all. That said, we think as a general matter that advertising does not fit into Category I. Some people find it annoying, and most people—we suspect—regard it as a cost of doing business with companies rather than an aim of the consumer in entering into the relationship.

Rather, advertising is perhaps the prototypical example of Category II, which is composed of data uses that advance the company's interests but

that neither advance nor undercut the consumer's interests. This category scores a win for the business but is value-neutral from the standpoint of the data's originator.

Along with advertising, a lot of the private sector's Big Data analytic work might come under Category II. Take an e-commerce site that scrutinizes a particular customer's historical purchasing habits and draws inferences about her interests or needs so as to market particular products in the future to her or to others, to sensibly establish discount percentages, or to set inventory levels in a more economical way. Or consider a cloud-based e-mail system that examines, on an automated and anonymized basis, the text of users' sent and received messages in an effort to better populate the ad spaces that border the area where users draft and read their e-mails.

Neither the online shopper nor the e-mail account holder obviously benefits from the above scenarios, but there isn't any measurable injury to speak of either. The consumer may like the ads, may find them annoying, or may look right through them and not care one way or the other. But in and of themselves, the ads neither advantage him nor do him harm.

Category II uses often bother some privacy activists.[14] In our view, however, it is better understood as a perfectly reasonable data-in-exchange-for-service arrangement. This is particularly true when Category II uses follow reasonably from the context of a consumers' interaction with a company.[15] People understand that targeted marketing is one of the reasons companies provide free services in exchange for consumer data, and they factor that reality into their decision to do business with those companies. As long as the companies are up front about what they are doing, this category of activity involves a set of judgments best regulated by consumer choice and preference.

This area is a good example of the tendency of privacy rhetoric to overpromise with respect to the protections consumers really need—or want. Seen through the lens of broader visions of privacy, a lot of Category II activity may cause anxieties about having data "out there" and about Big Data companies knowing a lot about us and having the ability to profile us and create digital dossiers on us.[16] But seen through a more modest databuse lens, these are relationships into which a reasonable consumer might responsibly choose to enter with reputable companies—indeed, they are choices that hundreds of millions of consumers are making every day worldwide. There is no particular reason to protect people preemptively from them.

Rather, database, in our view, can reasonably be defined as data uses in Category III; that is, those that run directly contrary to consumers' interests and either harm them discernibly, put them at serious and inherent risk of tangible harm, or run counter to past material representations made by the company to the consumer about things it would or would not do. Previously, we explained that databuse is the "right to not have your data rise up and attack you."[17] Category III includes data uses that advantage the corporate actor at the expense of the interests of the consumer. Category III activity should, in our view, provoke regulatory action and expose a company to injunction, civil penalty, or a money judgment—or even criminal prosecution, in the most egregious cases.

That makes Category III the most straightforward of our three-tiered scheme and examples of it far easier to identify. A company can be justly punished when it breaks a material promise made to the people who gave the company its data, such as: by using the data in a manner contradicted by a privacy policy or some other terms-establishing document; when the company stores its users' data in a less than reasonably safe way, such as by refusing to mitigate readily discoverable, significant cyber vulnerabilities or by failing to enact industry-standard and business-appropriate security practices; or when the company deploys data in a fashion that otherwise threatens or causes tangible injury to its customers.

The critical question for a corporation of any real size providing free services to customers and using their data is *how to keep a healthy distance from Category III activities while at the same time maximizing value.* The answer has to do with the trusteeship obligations businesses incur when they strive to make profitable use of their customers' data. These often imply a greater threshold of care and protection than purely market-oriented principles do. Trusteeship is normative in that it is designed to ensure a beneficiary's confidence and create conditions for the beneficiary's success. Market principles are ambivalent and thus suggest a just-do-the-least-required-to-further-one's-own-

ends sort of regime. A pure market approach would tolerate, for example, a minimally adequate corporate policy about how data is collected, used, and disseminated, or it would permit that policy to be scattered about various pages of a website, nested in a Russian-doll-like array of click-through submenus or drowned in legalese or technical gobbledygook. The good data trustee is going to do something more generous than that. Companies engaged in good data trusteeship will provide prominent, readily comprehensible explanations of their data practices, ones that fully equip the consumer to make informed choices about whether to do business or go elsewhere.[18]

The same idea holds true in other areas relevant to the two-sided arrangement between the person contributing data and the company holding data. The market might only require the company to obtain a consumer's consent to its data practices once. A good data trustee is going to refresh that consent regularly by giving the user a lot of control for so long as the user's data resides with the company. Where the market might presume consent to most uses generally, a good data trustee will not and instead will require additional consent for uses beyond those reasonably or necessarily following from the nature of the consumer's transaction with the company.[19]

It's easy to see what consumers get out of this vision, but what's in it for the companies? A lot. Trusteeship promises corporations the greatest possible measure of consumer confidence, and thus, a greater willingness to offer up more and more data for corporate use. As the FTC has reported, some of our economy's most data-deluged enterprises have found that the more choices they offer to their users in the first instance about whether to allow data exploitations, the more those users elect to remain "opted in" to features that use or disseminate data more broadly than the alternatives.[20] Getting people to give you large quantities of data requires, in the long run, their confidence. Good data trusteeship is critical to maintaining that confidence.

Conclusion

Consumers, governments, and companies need more guidance than the broad concept of privacy can meaningfully furnish. As a narrowing subset, databuse does a better job of portraying the government's current consumer protection efforts and legislative ambitions. In that respect, it offers all parties a firmer sense of what sorts of data uses actually are and are not off-limits. That's to the good, given that everyone wants greater clarity about the protections consumers actually require—and actually can expect—as against companies that ingest data constantly and by the boatload.

That's the difficulty with vague privacy talk—it disparages data usages by companies that don't measurably harm the companies' customers. The FTC doesn't sue companies simply because they stir fears, and the Commission really isn't asking for statutory power to do so either. Nor is the White House in its proposal for a Consumer Privacy Bill of Rights—which, again, largely recommends policies that jibe with our approach.

By observing that databuse better describes the government's behavior and short-term aspirations for consumer protection, we do not mean to proclaim the current setup to be optimal or to counsel against further legislation. To the

"GETTING PEOPLE TO GIVE YOU LARGE QUANTITIES OF DATA REQUIRES, IN THE LONG RUN, THEIR CONFIDENCE. GOOD DATA TRUSTEESHIP IS CRITICAL TO MAINTAINING THAT CONFIDENCE."

extent current law is not yet framed in terms of databuse—and it is not—the protections the FTC has quite reasonably grafted onto the unfair and deceptive trade practice prohibitions of the Federal Trade Commission Act should probably be fixed in statute. And if Congress wants to go further, it should raise standards too by more uniformly requiring the sorts of practices we hold out as models of good trusteeship.

But what the government should not do is push past databuse's conceptual boundaries and step into a more subjectively flavored, loosely defined privacy enforcement arena. We do not make law to defend "democracy" in its broadest sense; we subdivide that umbrella value into campaign finance law, redistricting, and other more manageably narrow ideas. The same holds true for "privacy," which, as a concept, is simply too gauzy, too disputed to serve as a practical guide. As its most fervent advocates understand it, it is a concept that might actually protect consumers far more than they wish to be protected. The costs of a sweeping "privacy" approach may well be to stifle and impede the delivery of services large numbers of people actually want. But isolating the core we actually mean to guarantee is one way of guaranteeing that core more rigorously.

ENDNOTES

1 Benjamin Wittes, "Databuse: Digital Privacy and the Mosaic," *The Brookings Institution*, 1 April 2011, 17.

2 Some companies have sought to offer customers a freestanding ability to make money from corporate uses of personal data—for example, by "giv[ing] users a cut of ad revenue." David Zax, "Is Personal Data the New Currency?" *MIT Technology Review*, 30 Nov. 2011, describing the now-defunct "Chime.In," a social networking site that split advertising sales with its members; *see also*, Joshua Brustein, "Start-Ups Seek to Help Users Put a Price on Their Personal Data," *The New York Times*, 12 Feb. 2013, describing early-stage efforts by companies to permit consumers to profit from data sales.

3 Restatement (Third) of Trusts § 78.

4 Ibid., § 77.

5 Ibid., § 82.

6 *See generally*, "Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy," *White House Report*, Feb. 2012.

7 *See*, e.g., "Consumer Data Privacy," 19, recommending consumer "right to secure and responsible handling of personal data"; "Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers at 22," *Final FTC Report*, 2012, 24-26, 24, recommending that companies "provide reasonable security for consumer data," noting recognition that the data security requirement is "well-settled"; "Commercial Data Privacy and Innovation in the Internet Economy: a Dynamic Policy Framework," *Department of Commerce Report*, 2010, 57, advocating for "comprehensive commercial data security breach framework."

8 *See,* "Consumer Data Privacy," 14, recommending consumer "right to easily understandable and accessible information about privacy and security practices"; "Protecting Consumer Privacy," viii, recommending, among other things, that companies "increase the transparency of their data practices," and that "privacy notices should be clearer, shorter, and more standardized to enable better comprehension and comparison of privacy practices"; "Commercial Data Privacy and Innovation," 30, arguing that information disclosed to consumers regarding companies' data practices should be "accessible, clear, meaningful, salient and comprehensible to its intended audience."

9 *See,* "Consumer Data Privacy," 11, recommending consumer "right to exercise control over what personal data companies collect from them and how they use it"; "Protecting Consumer Privacy," i, observing that recommendations of simplified choice and enhanced transparency would "giv[e] consumers greater control over the collection and use of their personal data"; "Commercial Data Privacy and Innovation," 69, "A key goal is to protect informed choice and to safeguard the ability of consumers to control access to personal information."

10 *See*, e.g., "Basic Privacy Settings & Tools" <https://www.facebook.com/help/325807937506242>; "Advertising on Facebook" <https://www.facebook.com/about/ads/#impact>.

ENDNOTES CONTINUED

11 *See*, e.g., "Ads Settings" <www.google.com/settings/ads>.

12 *See*, e.g., Complaint, In the Matter of Facebook, Inc., No. C-4365 ¶¶ 17-18 (July 27, 2012); Complaint, In the Matter of Myspace LLC, No. C-4369 ¶¶ 14-16, 21-28 (30 Aug. 2012).

13 This is but one application of the context principle, which the Obama administration has emphasized in its approach to consumer privacy. *See generally*, Helen Nissenbaum, "Privacy as Contextual Integrity," Washington Law Review, 79, no. 119 (2004); Helen Nissenbaum, *Privacy in Context: Technology, Policy and the Integrity of Social Life* (Stanford Law Books, 2010); *see also*, "Consumer Data Privacy," 15-19, advocating for consumers' right to expect that data collection and use will be handled in a manner consistent with the context in which consumers furnish their data; "Protecting Consumer Privacy," 36, stating that "[c]ompanies do not need to provide choice before collecting and using consumers' data for commonly accepted practices," including product fulfillment and fraud prevention; "Commercial Data Privacy and Innovation," 18 and n. 11, "A wide variety of authorities recognize that information privacy depends on context and that expectations of privacy in the commercial context evolve."

14 Jon Healey, "Privacy Advocates Attack Gmail – Again – for Email Scanning," *The Los Angeles Times*, 15 Aug. 2013, noting complaint by Consumer Watchdog, a consumer privacy organization, which challenged Google's scanning of messages sent to Google subscribers from non-Google subscribers; "Order Granting In Part and Denying In Part Defendant's Motion to Dismiss, In Re: Google Inc. Gmail Litigation," No. 13-MD-02340-LHK (N.D. Cal., Sept. 26, 2013), partially denying motion to dismiss where, among other things, plaintiffs alleged that Gmail's automated scanning protocols, as applied to inbound messages, had violated federal and state wiretapping laws.

15 See Footnote 13.

16 "Protecting Consumer Privacy in an Era of Rapid Change: a Proposed Framework for Businesses and Policymakers at 20," *Preliminary FTC Report*, 2012.

17 "Databuse," 4.

18 See Footnote 8.

19 See Footnote 13.

20 "Protecting Consumer Privacy," 9 and n. 40, noting, among other things, comments from Google regarding its subscribers, who use Google's Ads Preference Manager and remain "opted in."

LITERATURE SEARCH

A Reading List on the Data-Driven Economy

We can see the growth of our data-driven economy in the abundance of new thinking around its core concepts. That is why the U.S. Chamber of Commerce Foundation has compiled an annotated digital library of more than 400 documents broken down by theme to help you quickly understand the facets of our increasingly data-driven world. This compendium provides you with links to the best data thinkers out there.

FIND THE LIST AT

www.uschamberfoundation.org/reading-list-data-driven-economy

CONCLUDING THOUGHTS

THE ESSENTIAL INGREDIENT—US

BY RICH COOPER

Few issues have generated as much concern and confusion as the rise of Big Data, but given the impact on every industry, community, and person, should we expect anything less? Data is simply all of the information around us and about us. Yet, for all the reasons described in this report, it can be hard for individuals, policymakers, and other public and private sector stakeholders to fully comprehend and appreciate the data movement.

As a result of this incomplete knowledge and awareness, there is sometimes an unwarranted fear of data, a concern that information produced by the Internet of Things and all the technologies we use is reducing human beings to sets of 1s and 0s. As the age-old sentiment says, people fear what they do not understand. For many people, this fear boils down to a simple question: could the rise of data come to replace human intelligence?

In short, no. As the research and scholars in this report show, data does not work that way. Data contributes to informed decision making, but it is only a part of the equation. As history has proven time and again, being a good leader is about making good choices. In every setting, leaders must use a mix of reliable information and experience to decide the best course of action. The growing saturation of data-generating technologies contributes to an ocean of information that, when analyzed, can reveal new connections, trends, and opportunities. Yet, in the end, it will always be a person with a heartbeat (not an algorithm) that makes a final decision.

A recent report from The Economist Intelligence Unit, "Decisive Action: How Businesses Make Decisions and How They Could do it Better," investigated how intuition fits into business executives' decision-making processes.[1] In a survey of company leaders, the study found that 42% of respondents characterized their decision-making style as data-driven, while 17% noted a primarily empirical decision-making process. Just 10% reported a largely intuitive decision-making style.

"THE DATA AND THE DECISION-MAKER MUST WORK TOGETHER TO PRODUCE GROUNDBREAKING INNOVATIONS AND BUSINESS INSIGHTS."

Yet, when asked what they would do if data contradicted a "gut feeling," nearly 60% of business leaders said they would reanalyze the data; 30% said they would collect more data; and, a meager 10% would ignore that little voice inside and do what the data says.

What this tells us is that even as data-driven decision making is an important and growing force, it does not trump good, old-fashioned human intuition. Nor should it. For all of the powerful, valuable insights data can offer, it can never replace a conversation between parties, an experience-based deduction, or any of the un-replicable cognitive qualities unique to human beings.

Data veracity is a challenge for analysts. This refers to data accuracy as well as source reliability, the context out of which the data comes, the methods for sorting and storing information, and a range of factors that can influence the data's validity. Remedying this is already a large, time-consuming effort. The *Harvard Business Review* reports that workers can spend up to 50% of their time looking for data, fixing errors, and trying to validate the numbers they have on hand.[2]

While this shows the ongoing challenge of acquiring high-quality data, it also underscores another way in which the human element remains critical. Collecting data and preparing it for analysis still demands a human intelligence. It is that intuitive hunch, that gut feeling that can push a business leader to pause before acting on data analysis that just doesn't add up. Without human knowledge and wisdom, we might end up chasing Big Data red herrings. Instead, data informs our thoughts, actions, and discussions and elevates them to a higher level. The data and the decision-maker must work together to produce groundbreaking innovations and business insights.

Balancing Needs and Opportunities

Each of the chapters in this report discusses the core, interdependent attributes of the data-driven world. Realizing the most value from data is a careful balancing act, with multiple competing priorities that must receive appropriate attention and commitment or we will not enjoy the jobs, innovations, efficiencies, and better quality of life that data can yield.

As Leslie Bradshaw writes in Chapter 3, the data movement is akin to the era-advancing technological breakthroughs of centuries past, which included the printing press, the steam engine, and the semiconductor. Like Big Data, each of these technologies presented a steep learning curve for society, demanding knowledge for effective application. Recognizing that the human element is an indispensable part of the data movement, the challenge for modern society is to foster data literacy among policymakers, business leaders and entrepreneurs, and citizens, such that we can realize data's value.

This value is substantial. Data brings enormous opportunities for growth. It drives innovation and business success, which in turn deliver cascading economic and productivity gains. Indeed, as McKinsey finds (and as Joseph Kennedy cites in

Chapter 2), the better use of data could increase global income by $3 trillion each year in just seven industries. These potential economic gains are compounded if data is shared between organizations. For example, in 2012, the data-driven marketing economy topped $156 billion and created 676,000 jobs—70% of this value and employment depended on moving data between firms.[3]

While dollar figures are important, so too are the competitive benefits that come with data-driven business and innovation. John Raidt writes in Chapter 4 that the United States is unrivaled in its capacity to extract the most value from data. Yet, other countries are also looking at how data can help them, and the United States must continue fostering the data-driven economy with targeted steps towards greater competitiveness. This includes building a vibrant and dynamic STEM workforce, expanding a robust broadband infrastructure, developing trade agreements and practices that ensure the flow and protection of data, and adjusting publicly funded R&D to better develop data capabilities and public-private collaboration.

One way collaboration can be encouraged is through the principle of Open Data. As Joel Gurin describes in Chapter 6, Open Data can be used by anyone as a free (or low-cost) public resource and can be used to start new businesses, gain business intelligence, and improve business processes. While Open Data is not limited to public sector data, the most extensive, widely used Open Data comes from government agencies and offices. As such, governments at all levels need to develop policies and processes to release relevant, accessible, and useful Open Data sources to enable innovation, support a better-informed public, and create economic opportunity. By doing so, we unleash untapped potential in our economy and workforce, providing benefits and linking entrepreneurs, consumers, and average citizens in every region of the country.

All of this demands a national strategic plan for properly aligning public policies, resources, and priorities. As Matthew Harding writes in Chapter 5, Big Data needs to be grounded on open standards and requires advanced technological solutions to monitor and enforce high quality in acquisition and use. Clear principles of data ownership are urgently required. Public policies should encourage responsible use of data. Privacy and security concerns are best addressed by industry-led initiatives that are flexible, innovative, and technologically sound. Policies that restrict or prevent data access and sharing are a major impediment to innovation and public welfare.

Effective, strategic policies are just as important in the private sector, as corporations and governments alike face complex questions about data ownership and use. These questions are frequently (and unfortunately) reduced to vague calls for privacy, but as Benjamin Wittes and Wells C. Bennett note in Chapter 7, privacy is actually not an accurate word for discussing corporate responsibilities and consumer protection. The word promises a great deal more than policymakers are prepared to deliver, and in some ways, it also promises more than consumers want. Rather, what is needed is protection against "databuse"—the malicious, reckless, negligent, or unjustified handling, collection, or use of a person's data in a fashion adverse to that person's interests and in the absence of that person's knowing consent. Companies must be reasonable and honest data custodians, handling data in a forthright and secure fashion, one that does not injure consumers and gives them reasonable information about and control over what is being done with their data.

Towards a New World

The data movement is unlike anything we have seen before. It connects all people, activities, and the goods and services we create. It transcends national borders and arbitrary barriers between people, cultures, and ideas. This new world, shaped by data, gives us a rare opportunity to explore and discover.

An historic parallel are the adventures of the first nautical explorers. As these early, ultimate risk-takers looked out from the shoreline, preparing to shove off into open waters, they had their sails at the ready to go somewhere, but where they would land and the direction they would take was often unknown. To be sure, the forces of nature certainly impacted those journeys, but it was human hands, firmly grasping the rudders of available technologies and innovations, that steered these pioneers to new shores of opportunity.

Today, we find ourselves on the verge of a similarly epic journey, its endpoint ever-uncertain. The direction we take will be decided by the forces of

nature and commerce, as well as the debates and discussions we have about what the data-driven future should look like. The rudder by which we steer is held with imperfect hands, where error and discovery are just a few degrees apart. Yet, this is a journey we must take if we are to keep moving forward.

The winds of innovation are blowing, our sails are raised, and there is an ocean of data and possibility before us. At the outset of any adventure, a measure of anxiety can be healthy and helpful, but this ship of opportunity is of our making and fully within our control. If there is one lesson we can take with us on this voyage, it is this: the power of data is not what it can do but what we can do with it.

ENDNOTES

1 "Decisive Action: How Businesses Make Decisions and How They Could Do it Better," *The Economist Intelligence Unit*, June 2014.

2 Thomas C. Redman, "Data's Credibility Problem," *Harvard Business Review*, Dec. 2013.

3 John Deighton and Peter A. Johnson, "The Value of Data: Consequences for Insight, Innovation and Efficiency in the U.S. Economy," *Data-Driven Marketing Institute*, 14 Oct. 2013.

ABOUT THE AUTHOR

Rich Cooper is vice president for emerging issues and research for the U.S. Chamber of Commerce Foundation where he is responsible for the exploration of issues that will impact the private sector over the next three to five years. He directs a team of scholars, researchers, and managers who present programming, publications, and events to better inform and best prepare business leaders for the future. He is a senior fellow with The George Washington University's Homeland Security Policy Institute, the past chairman of the Homeland Security Division of the National Defense Industrial Association, and has previously held senior positions at the U.S. Department of Homeland Security, NASA, and in the private sector.

ACKNOWLEDGEMENTS

The U. S. Chamber of Commerce Foundation would like to gratefully acknowledge the contributions, insights, and energies of a number of people who helped inform this work on the future of data-driven innovation.

U.S. Chamber of Commerce Foundation
The Honorable John R. McKernan
Al Martinez-Fonts
Rich Cooper
Michael Hendrix
Dr. Jeff Lundy
Tim Lemke
Melissa Guay
Tony Mills
John Drugan

U.S. Chamber of Commerce
David Chavern
Rebecca Oliver
Jason Goldman
Adam Schlosser
Bradley Hayes
Matthew Eggers
Brian Noyes
Frank Cullen
Andrew Kovalcin
Jess Sharp
Brian Miller

Contributing Researchers
Dr. Joseph Kennedy, *Kennedy Research LLC*
Joel Gurin, *The GovLab*
Dr. Matthew Harding, *Duke University*
Leslie Bradshaw, *Made by Many*
John Raidt, *Atlantic Council & Jones Group International*
Benjamin Wittes, *Brookings Institution*
Wells C. Bennett, *Brookings Institution*
Rich Cooper, *U.S. Chamber of Commerce Foundation*

Copy Editor & Report Consultant
Justin Hienz, Cogent Writing, LLC

Design & Illustration
Beutler Ink

U.S. CHAMBER OF COMMERCE FOUNDATION

1615 H ST. NW | WASHINGTON, DC 20062-2000
WWW.USCHAMBERFOUNDATION.ORG