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.
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. 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.”

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. 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. 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. 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 misrecorded 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.”

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.

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. 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. 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.
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.\textsuperscript{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.\textsuperscript{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 customer 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:\textsuperscript{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
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. 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. 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. 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.

“We need new industry-driven solutions that are flexible enough to promote the responsible use of individual customer data.”
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


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


17 Calo, “Consumer Subject Review Boards.”

18 “Coding in schools”, The Economist, 26 April 2014.
