TECHNOLOGY POLICY INSTITUTE

Studying the Global Information Economy

Big Data and Consumer Privacy in the Internet Economy
Comments of Thomas M. Lenard, PhD, President and Senior Fellow
Technology Policy Institute

Submitted to the National Telecommunications and Information Administration
U.S. Department of Commerce
August 5, 2014

INTRODUCTION

These comments respond to the June 6, 2014 National Telecommunications and Information Administration (NTIA) request for comments on privacy policy issues raised by two May 2014 White House reports on big data—one by a team led by Presidential Counselor John Podesta (the Big Data Report), and a complementary study by the President's Council of Advisors on Science and Technology (the PCAST Report).[1]

In light of these two reports, the NTIA notice raises the question of "how big data might impact the protections called for in the Consumer Privacy Bill of Rights,"[2] as defined in an earlier (2012) White House report.[3] I address this issue, drawing from a 2014 Technology Policy Institute paper (attached).[4] I find that the recent White House big data reports raise serious

[1] Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values,* May, 2014 (Big Data Report); President's Council of Advisors on Science and Technology, *Report to the President, Big Data and Privacy: A Technological Perspective,* May, 2014 (PCAST Report).
[2] http://www.ntia.doc.gov/press-release/2014/ntia-seeks-comment-big-data-and-consumer-privacy-bill-rights
[3] The White House, *Consumer Data Privacy in a Networked World: A Framework For Protecting Privacy and Promoting Innovation in the Global Digital Economy*, February, 2012 (White House 2012 Report).
[4] Lenard, Thomas M., and Rubin, Paul H. "Big Data, Privacy and the Familiar Solutions" *Technology Policy Institute*, May 2014. http://www.techpolicyinstitute.org/files/lenard_big%20dataprivacyandthefamiliarsolutions.pdf

questions about the principles of limiting data collection and use, access, and others on which the Consumer Privacy Bill of Rights (CPBR) is based. The White House reports suggest that privacy policy should instead focus on uses of data that cause actual consumer harm.

IS THERE EVIDENCE OF HARM?

The threshold question in determining whether to adopt the CPBR or any privacy regulation at all is whether there is evidence of a market failure or consumer harm. Neither of the two recent White House reports documents such evidence. Indeed, the government has in the past two years issued five reports—including the two most recent—that fail to present evidence that data (big or small) used for commercial and other non-surveillance purposes have caused actual privacy harms.[5] Discussions of harm in these reports are hypothetical and speculative.

The NTIA notice mentions two specific areas of concern: data brokers, and the potential for big data analysis to discriminate against particular groups. Yet, the recent reports do not contain evidence justifying those concerns.

The Federal Trade Commission recently released a detailed report on the operations of data brokers, describing in detail many benefits, but presenting no evidence of actual harms.[6] Instead, the Commission points to "potential risks" from data brokers, such as advertising that "some consumers may find troubling," and marketing classifications that "may be disconcerting."

[5] The Big Data Report; the PCAST Report; the White House 2012 Report; the Federal Trade Commission, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers*, March, 2012; and the Federal Trade Commission, *Data Brokers: A Call for Transparency and Accountability*, May, 2014 (Data Broker Report).

[6] Data Broker Report

Writers who argue that data collection and analytics favor the rich over the poor rely on hypothetical rather than actual examples.[7] The Big Data Report argues that "big data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace,"[8] but again the supporting evidence is weak. Indeed, the two specific examples of discrimination cited in the Big Data Report turn out to be essentially non-examples. Also, both examples involve government—not private sector—use of data.

The first example involves StreetBump, a mobile application developed to collect information about potholes and other road conditions in Boston. Even before its launch, the city recognized that this app, by itself, would be biased toward identifying problems in wealthier neighborhoods, because wealthier individuals would be more likely to own smartphones and make use of the app. As a result, the city adjusted accordingly to assure reporting of road conditions was accurate and consistent throughout the city.[9]

The second example involves the E-verify program used by employers to check the eligibility of employees to work legally in the United States. The report cites a study that "found the rate at which U.S. citizen (sic) have their authorization to work be initially erroneously unconfirmed by the system was 0.3 percent, compared to 2.1 percent for non-citizens. However, after a few days many of these workers' status was confirmed."[10] It seems almost inevitable that the error rate for citizens would be lower, because citizens are automatically eligible to work, whereas additional

[7] See, for example, Joseph Jerome, "Buying and Selling Privacy: Big Data's Different Burdens and Benefits," 66 Stanford Law Review Online 47, 2013; and Omar Tene, "Privacy: For the Rich or for the Poor," Concurring Opinions Blog Post, June 2013, available at: http://www.concurringopinions.com/archives/2012/07/privacy-for-the-rich-or-for-the-poor.html.
[8] Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values,* May, 2014, Cover Letter.
[9] Big Data Report, pp. 51-52.
[10] Big Data Report, p. 52.

information is needed to confirm eligibility for non-citizens (i.e., evidence of some sort of work permit). Hence, it is not clear this is an example of discrimination.

If the concern is about the use of big data to facilitate price discrimination, it is more likely that price discrimination will favor lower-income individuals. Since price discrimination involves charging different prices to different consumers for the same product based on their willingness to pay, and since willingness to pay is generally positively related to ability to pay, price discrimination will, other things equal, result in lower prices to lower-income consumers.

Moreover, big data are being used to develop products that specifically benefit lower-income consumers. For example, companies, such as ZestFinance, LendUp, and Better Finance Inc. are using models containing many more variables than traditional credit scoring to help lenders determine whether or not to offer small, short-term loans to people who are otherwise poor credit risks.[11] This provides a better alternative to people who otherwise might rely on payday lenders or even loan sharks. Other companies, such as Kabbage and On Deck Capital, provide lending services to very small businesses.[12]

BIG DATA AND THE CONSUMER PRIVACY BILL OF RIGHTS

The CPBR is based on the Fair Information Practice Principles (FIPPs), which consist of notice, choice, access and security. These principles, which date back to the 1970s, are reflected in the OECD Privacy Principles of 1980, current European Union regulations, and the recommendations of the FTC's 2012 Privacy Report, and have been the focus of privacy policy

[11] ZestFinance see http://www.zestfinance.com/how-we-do-it.html. LendUp see https://www.lendup.com/about. Better Finance Inc. see http://www.smartpaylease.com/.

[12] Kabbage see https://www.kabbage.com/how-it-works. OnDeck see https://www.ondeck.com/company/

discussions for several decades.[13] The central thrust of these policies is to limit the collection and use of information. The NTIA notice asks how these principles work in a world of big data.

As the NTIA notes, "Some privacy experts believe nuanced articulations of these principles [the FIPPs] are flexible enough to address and support new and emerging uses of data, including big data. Others, especially technologists, are less sure, as it is undeniable that big data challenges several of the key assumptions that underpin current privacy frameworks, especially around collection and use."

Both of the recent White House reports indicate that the focus on limiting data collection is increasingly irrelevant and, indeed, harmful in a big data world. The Big Data Report observes that "these trends may require us to look closely at the notice and consent framework that has been a central pillar of how privacy practices have been organized for more than four decades."[14] The PCAST Report notes, "The beneficial uses of near-ubiquitous data collection are large, and they fuel an increasingly important set of economic activities. Taken together, these considerations suggest that a policy focus on limiting data collection will not be a broadly applicable or scalable strategy—nor one likely to achieve the right balance between beneficial results and unintended negative consequences (such as inhibiting economic growth)."[15]

This is primarily because big data analysis typically involves uses of data that were not anticipated at the time the data were collected. It also frequently involves the combination of data from different sources.

---

[13] An excellent summary of the evolution of the FIPPs comes from Robert Gellman, "FAIR INFORMATION PRACTICES: A Basic History", last updated November 11, 2013, available at http://www.bobgellman.com/rg-docs/rg-FIPShistory.pdf, and the current FTC FIPPs are posted at http://www.ftc.gov/reports/privacy3/fairinfo.shtm.

[14] Big Data Report, p. 54.

[15] PCAST Report, pp. x-xi.

Indeed, a major implication of the CPBR/FIPPs framework is that data should only be collected for a specific, identified purpose. But limiting the reuse or sharing of data is inconsistent with the innovative ways in which big data are being used. Such limitations would adversely affect innovation in a variety of sectors, from marketing to credit markets to health research.

Principles of notice and choice become almost meaningless when data may be used in unpredictable ways. As the PCAST report notes, "As a useful policy tool, notice and consent is defeated by exactly the positive benefits that big data enables: new, non-obvious, unexpectedly powerful uses of data. It is simply too complicated for the individual to make fine-grained choices for every new situation or app."[16] Moreover, "Only in some fantasy world do users actually read these notices and understand their implications before clicking to indicate their consent."[17]

Another element of the CPBR/FIPPs framework is transparency.[18] The notion that consumers should understand who is collecting their data and how they are being used is an appealing one, but it is largely meaningless in the big data era where scores may be based on hundreds of data points and complex calculations. This information cannot be meaningfully conveyed through a notice, and consumers would not devote the hours required to understand it. For example, it is not clear that a person rejected for credit by a complex algorithm would particularly benefit by being shown the equation used. The FICO score, an early example of a calculation based on a

[16] PCAST Report, p. 38.
[17] PCAST Report, p. xi.
[18] The White House, Consumer Data Privacy in a Networked World: A Framework For Protecting Privacy and Promoting Innovation in the Global Digital Economy, February, 2012 pg. 14-15

complex algorithm, is virtually impossible to explain to even an informed consumer because of interactions and nonlinearities in the way various data points enter into the score.[19]

The CPBR/FIPPs framework also advocates data "access and accuracy." Giving consumers the ability to correct their information may be more complicated than it might appear, even aside from the administrative complexities. For example, consumers do have the right to correct information used in deriving their credit scores, but it is made difficult to do so, for good reason. An individual who thinks she has been wrongly categorized clearly has an interest in correcting erroneous information if that information has a negative effect. But she might also have an interest in "correcting" valid information that would adversely affect the decision, or inserting incorrect information that would have a positive effect. Distinguishing between these various "corrections" may be quite difficult.

The purpose of collecting information that affects decisions about individuals—e.g., credit decisions, insurance decisions, or employment decisions—is to ameliorate an asymmetric information problem. Individuals have much more information about themselves than lenders, insurance companies or prospective employers. Asymmetric information is a feature of some markets that potentially can lead to market breakdown.[20] This is why, as former FTC officials Howard Beales and Timothy Muris point out, "In our economy, there are vital uses of information sharing [such as credit reporting] that depend on the fact that consumers cannot choose whether to participate."[21]

---

[19] The major inputs to a credit score are well known; however, the calculation of credit scores from credit report data is proprietary and exceedingly complex. See for example FDIC, *Credit Card Activities Manual*, Ch. 8 – Scoring and Modeling, 2007, available at: http://www.fdic.gov/regulations/examinations/credit_card/.

[20] See George A. Akerlof, "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism", *The Quarterly Journal of Economics,* Vol. 84, No. 3, August, 1970, pp. 488-500.

[21] Beales, J. H., and Muris, T. J. "Choice or consequences: Protecting privacy in commercial information." *The University of Chicago Law Review*, p. 115.

Moreover, if we make it easier for individuals to access their data then we also make it easier for those bent on fraud to access the same data. If fraudsters have access to large amounts of data about a person, they can more easily defraud that individual (perhaps by making purchases that are consistent with the individual's behavior in order to trick the credit card companies' monitoring efforts). Thus, ease of consumer monitoring of data is at best a two-edged sword.

AN ALTERNATIVE APPROACH

As an alternative to the standard CPBR/FIPPs approach, the Big Data Report suggests examining "whether a greater focus on how data is used and reused would be a more productive basis for managing privacy rights in a big data environment."[22] The PCAST Report is even clearer:

> Policy attention should focus more on the actual uses of big data and less on its collection and analysis. By actual uses, we mean the specific events where something happens that can cause an adverse consequence or harm to an individual or class of individuals…. By contrast, PCAST judges that policies focused on the regulation of data collection, storage, retention, a priori limitations on applications, and analysis… are unlikely to yield effective strategies for improving privacy. Such policies would be unlikely to be scalable over time, or to be enforceable by other than severe and economically damaging measures.[23]

Beales and Muris also recommend an approach based on "the consequences of information use and misuse. There is little basis for concern among most consumers or policymakers about information sharing per se. There is legitimate concern, however, that some recipients of the

[22] Big Data Report, p. 61.
[23] PCAST Report, p. xiii.

information will use it to create adverse consequences for the consumer."[24] As an example of a consequence-based policy, they point to the FTC's Do Not Call Registry, aimed at the adverse consequence of receiving unwanted marketing calls.

CONCLUSION

Both the Big Data and PCAST Reports are clear that "[t]he beneficial uses of near-ubiquitous data collection are large, and they fuel an increasingly important set of economic activities."[25] This implies that limiting the collection and use of data in a way that is not addressed at specific harms would likely be very costly.

Thus, the conclusions of the White House reports are inconsistent with the CPBR/FIPPs approach. These reports can serve a very useful purpose if they refocus privacy discussions where they should be focused—on actual harms to individuals. Up to now, the various government and academic studies have not found evidence of harms from the use of data for commercial and other non-surveillance purposes. If evidence of harm is found, the next question for policy makers would be whether there is a remedy available that reasonably can be expected to yield benefits greater than costs.

[24] Beales and Muris, p. 118
[25] PCAST Report, p. x.

TECHNOLOGY POLICY INSTITUTE
Studying the Global Information Economy

Big Data, Privacy and the Familiar Solutions*

May 2014

Thomas M. Lenard and Paul H. Rubin

* Prepared for Public Policy Conference on "The Future of Privacy and Data Security Regulation," George Mason University Law School, Law and Economics Center, May 14, 2014.

BIG DATA, PRIVACY AND THE FAMILIAR SOLUTIONS

Thomas M. Lenard and Paul H. Rubin*

I. Introduction

The information technology revolution has produced a data revolution—sometimes referred to as "big data"—in which massive amounts of data are collected, stored and analyzed at relatively low cost. An integral part of the big data revolution is the rapidly developing Internet of Things (IoT), also known as the Internet of Everything, which generates a growing supply of devices and objects from which data can be gathered.

The emergence of big data and the IoT has raised concerns on the part of some privacy scholars, advocates and government officials.[1] Federal Trade Commission Chairwoman Edith Ramirez devoted her first major speech on privacy to big data, arguing that "the challenges [big data] poses to privacy are familiar, even though they may be of a magnitude we have yet to see."[2] She added, "The solutions are also familiar, [a]nd, with the advent of big data, they are now more important than ever." Chairwoman Ramirez's speech raised the question of whether big data are associated with new privacy harms and a concomitant increase in the need for government action. It also suggested that we should look to the standard solutions involving notice and choice, use specification and limits, data minimization and transparency to solve potential privacy problems brought about by big data.

Both the White House and the FTC completed major privacy reports in 2012.[3] Although neither report explicitly mentions big data or the IoT, their policy recommendations clearly would have a large impact on both.

The 2012 FTC report's principal recommendation of Privacy by Design (PBD) requires companies to "promote consumer privacy throughout their organizations and at every stage of the development of their products and services."[4] Substantively, PBD is essentially a restatement of the traditional Fair Information Practice Principles (FIPPs) of Notice, Choice, Access and Security: "The framework embodies all the concepts in the 1980 OECD privacy

---

* Lenard is President and Senior Fellow at the Technology Policy Institute. Rubin is Samuel Candler Dobbs Professor of economics at Emory University and Senior Fellow at TPI. The authors thank Arlene Holen, Amy Smorodin and Scott Wallsten for helpful comments, and Corwin Rhyan for outstanding research assistance.

[1] Much of the concern relates to the collection and use of data by governments for national security purposes, a subject of intense debate across the globe following the leaks by National Security Agency contractor Edward Snowden. This is obviously a major issue, but not the subject of this paper.

[2] Edith Ramirez, "The Privacy Challenges of Big Data: A View from the Lifeguard's Chair", Speech at Technology Policy Institute's Aspen Forum, August, 2013, accessed at http://ftc.gov/speeches/ramirez.shtm.

[3] The White House, *Consumer Data Privacy in a Networked World: A Framework For Protecting Privacy and Promoting Innovation in the Global Digital Economy*, February, 2012; The Federal Trade Commission, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers,* March, 2012.

[4] FTC Report, p. 22.

guidelines…. For example, privacy by design includes the collection limitation, data quality, and security principles. Additionally, the framework's simplified choice and transparency components… encompass the OECD principles of purpose specification, use limitation, individual participation, and openness."[5] The PBD framework "calls on companies to (1) delete consumer data that they no longer need and (2) allow consumers to access their data and in appropriate cases suppress or delete it." Finally, "Reasonable collection limits and data disposal policies work in tandem with streamlined notices and improved consumer choice mechanisms."[6]

In May 2014, the White House released two reports specifically focusing on big data, as a result of a 90-day study President Obama announced on January 17—one by a team led by Presidential Counselor John Podesta (the Executive Office of the President or "EOP" report), and a complementary study by the President's Council of Advisors on Science and Technology (the PCAST report).[7] The reports recognize big data's benefits and potential and suggest, in light of the way big data are used, a refocus of the policy discussion.

This paper proceeds as follows: In Section II, we discuss the promise of big data and present examples of their use in both the public and private sector. The examples show how big data provide the opportunity for significant innovation and value creation.

Section III focuses on potential privacy and security threats that have been highlighted by privacy advocates, scholars and public officials. Specifically, we address the following questions:

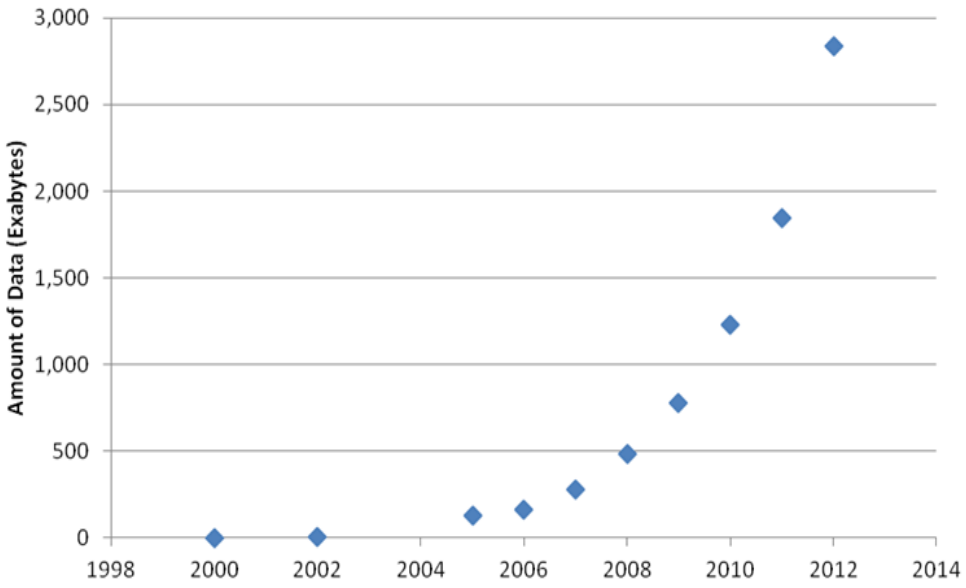
- What are the implications of big data for data security—data breaches and identity fraud?
- What are the implications of big data for profiling individuals and using algorithms to draw inferences for purposes ranging from marketing to credit and employment decisions?
- Does the use of big data introduce biases that can be considered discriminatory?
- Are firms using big data to manipulate consumers into buying goods or services they don't really want or are not beneficial?
- Does targeting and customization result in harm to consumers from a reduction in the variety of information to which consumers are exposed?
- Will big data force individuals to reveal too much information about themselves, thereby eroding privacy?

Section IV discusses policy proposals regarding big data. We first discuss whether there are identifiable harms attributable to big data that could be alleviated by government policies. We next analyze the standard solutions reflected in PBD, the FIPPs and the OECD principles in the context of big data. For example, we explore:

---

[5] FTC Report, p. 23.
[6] FTC Report, p. 24.
[7] Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values*, May, 2014 (EOP Report); President's Council of Advisors on Science and Technology, *Report to the President, Big Data and Privacy: A Technological Perspective*, May, 2014 (PCAST Report).

Figure 1: Digital Data Created Annually Worldwide



Source: *Digital Universe Reports*, IDC

- How should we think about the "reuse" of data—i.e., the use of data for purposes not initially identified or even envisioned?
- Similarly, how should we think about combining data from different sources?
- What are the implications of greater transparency about how data are being used?

Finally, we discuss alternative approaches that have recently been suggested by the White House reports and others, including targeting policies to specific misuses.

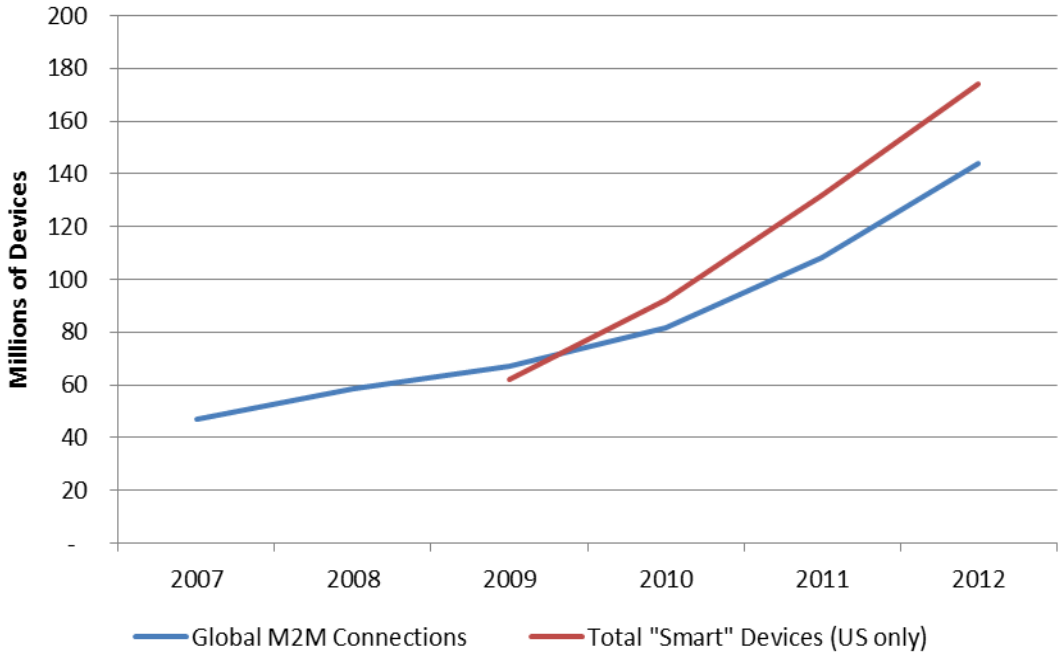
We conclude that there is no evidence at present that big data used for commercial and other non-surveillance purposes have caused privacy harms. Moreover, the standard solutions associated with PBD, the FIPPs and the OECD principles would impose barriers to the innovation expected from the big data revolution.

II. The Promise of Big Data

By all accounts, the use of data is increasing dramatically. One measure of the big data revolution is the increased flow of digital data, which has grown from an estimated 0.6 to 2.1 exabytes in 2000 to 2,700 exabytes in 2012, as shown in Figure 1. About one-third of the data collected globally is estimated to originate in the United States.[8]

[8] IDC, sponsored by EMC, *The Digital Universe in 2020: Big Data, Bigger Data Shadows, and Biggest Growth in the Far East*, December, 2012.

Figure 2: Global M2M Connections and US "Smart" Devices



Sources: *Global M2M Reports*, Berg Insight, Pyramid Research, & ABI Research; *Semi-Annual Wireless Survey Top Line Results*, CTIA

Mirroring the growth in data is the increase in the number of devices that might be considered part of the IoT. Cisco estimates that the number of connected devices worldwide grew from 500 million in 2003 to 12.5 billion in 2010 and will reach 50 billion by 2020.[9] Gartner projects that by 2020 there will be "only" 26 billion "units installed."[10] Machine-to-machine connections—connections that do not have a human interface—also measure the growth of IoT and have more than tripled between 2005 and 2012, as shown in Figure 2. Finally, Figure 2 also shows the growth of smart devices, which reflects the growth of the IoT.

While one may be skeptical of the hype surrounding big data, they clearly create the potential for significant innovation not only in specific sectors, but also in the overall economy. The 2014 EOP report starts out by observing that "properly implemented, big data will become an historic driver of progress."[11] Reports from the World Economic Forum, McKinsey Global Institute and others describe the potential benefits in such sectors as health care, government services, fraud protection, retailing, and manufacturing. McKinsey estimates that big data and analytics could

[9] Cisco Internet Business Solutions Group, *The Internet of Things: How the Next Evolution of the Internet Is Changing Everything*, April, 2011.
[10] Gartner, *Forecast: The Internet of Things, Worldwide*, 2013.
[11] Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values*, May 2014, Cover Letter.

yield benefits for health care alone of more than $300 billion annually. For the overall economy, gains could potentially be up to $610 billion in annual productivity and cost savings.[12]

Michael Mandel estimates that the Internet of Everything has the potential to increase GDP by 0.2 to 0.4 percentage points over the next 10-15 years:[13] "The Internet of Everything is about building up a new infrastructure that combines ubiquitous sensors and wireless connectivity in order to greatly expand the data collected about physical and economic activities; expanding 'big data' processing capabilities to make sense of all that new data; providing better ways for people to access that data in real-time; and creating new frameworks for real-time collaboration both within and across organizations."[14]

Although the term is now ubiquitous, "There is no rigorous definition of big data."[15] McKinsey defines big data as referring to "datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze."[16] Mayer-Schönberger and Cukier, in their recent book on big data, focus on what the data can produce: "Big data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value, in ways that change markets, organizations, the relationship between citizens and governments, and more."[17] They focus on the ability of large data sets to yield correlations between variables that can provide important public and private benefits.

Einav and Levin echo this point in a recent paper discussing the potentially revolutionary effects of data on economic analysis and policymaking. Big data's potential comes from "the identification of novel patterns in behavior or activity, and the development of predictive models, that would have been hard or impossible with smaller samples, fewer variables, or more aggregation."[18] Data are now available in real time, at larger scale, with less structure, and on different types of variables than previously.[19]

Innovative Uses of Big Data

The poster child for big data is Google Flu. Testing 450 million models, researchers identified 45 search terms that predict the spread of flu more rapidly than the Centers for Disease Control, which relies on physicians' reports.[20] By tracking the rate at which the public searches for terms

---

[12] McKinsey Global Institute, *Game changers: Five opportunities for US growth and renewal,* July, 2013, p. 66.
[13] Michael Mandel, *Can the Internet of Everything Bring Back the High-Growth Economy,* Progressive Policy Institute, September, 2013, p. 2.
[14] Mandel, pp. 2-3.
[15] Mayer-Schönberger and Cukier, "Big Data: a revolution that will transform how we live, work and think", Houghton Mifflin Harcourt, 2013, p. 6.
[16] McKinsey Global Institute, *Big Data: The Next Frontier for Innovation, Competition and Productivity,* June, 2011 p. 1.
[17] Mayer-Schönberger and Cukier, p. 6.
[18] Liran Einav and Jonathan Levin, "The Data Revolution and Economic Analysis", Prepared for the NBER Innovation Policy and the Economy Conference, April, 2013, p. 2.
[19] Einav and Levin, pp. 5-6.
[20] Mayer-Schönberger and Cukier, pp. 2-3. While Google Flu has generally been very accurate, there have been glitches. Google Flu seems to have overestimated the incidence of flu early in the 2013 season, because widespread

like "flu" and "cough medicine" using Google, an outbreak of influenza can be spotted a week or two ahead of CDC reports.[21]

Because big data analysis—as exemplified by Google Flu—involves finding correlations and patterns that might otherwise not be observable, it typically involves uses of data that were not anticipated at the time the data were collected. Mayer-Schönberger and Cukier emphasize that "in a big-data age, most innovative secondary uses haven't been imagined when the data is first collected." They add, "With big data, the value of information no longer resides solely in its primary purpose. As we've argued, it is now in secondary uses."[22]

Serendipitous uses of data are not, however, a new phenomenon or confined to the digital era. Mayer-Schönberger and Cukier give the example of Commander Mathew Maury, who, in the middle of the 19th century, used data from logbooks of past voyages to devise more efficient routes and mapped out the shipping lanes that are still in use today. His data were also used to lay the first transatlantic telegraph cable.[23] Commander Maury "took information generated for one purpose and converted it into something else."[24]

More recent examples of the unanticipated use of data are numerous. In the health care area, for example, the Danish Cancer Society combined Denmark's national registry of cancer patients with cell phone subscriber data to study whether cell phone use increased the risk of cancer.[25] The Food and Drug Administration used Kaiser Permanente's database of 1.4 million patients to show that the arthritis drug Vioxx increased the risk of heart attacks and strokes.[26] The CDC combine airline records, disease reports, and demographic data to track epidemics and other health risks.[27]

Einav and Levin survey new research by economists using large-scale, real-time data to better track and forecast economic activity using measures that supplement official government statistics. The Billion Prices Project, for example, uses data on retail transactions from hundreds of online retail websites to produce alternative price indices that are made available in real time, before the official Consumer Price Indexes.[28] In the same vein, Choi and Varian have used

---

press reports of a particularly severe outbreak may have induced more searches by people who didn't actually have the flu. See http://www.nature.com/news/when-google-got-flu-wrong-1.12413.

[21] Jeremy Ginsburg et al., "Detecting Influenza Epidemics Using Search Engine Query Data," *Nature*, Vol. 457, February 2009, pp. 1012-14, http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html.

[22] Mayer-Schönberger and Cukier, p. 153.

[23] Mayer-Schönberger and Cukier, pp. 73-76.

[24] Mayer-Schönberger and Cukier, p. 76.

[25] See the Danish study by Cardis et al, "The INTERPHONE study: design, epidemiological methods, and description of the study population", *European Journal of Epidemiology*, Vol. 22, No. 9, 2007, pp. 647-664.

[26] See the original study by Graham et al, "Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study", *The Lancet*, Vol. 365, No. 9458, 2005, pp. 475-481.

[27] A discussion of the new CDC tool, Biomosiac, can be found in Amy O'Leary, "In New Tools to Combat Epidemics, the Key is Context", *The New York Times Bits Blog*, June 19, 2013, accessible at http://bits.blogs.nytimes.com/2013/06/19/in-new-tools-to-combat-epidemics-the-key-is-context/?smid=tw-share.

[28] See http://bpp.mit.edu/usa/.

Google search engine data to provide accurate measures of unemployment and consumer confidence.[29] Wu and Brynjolfsson have used search data to predict housing market trends.[30]

In the private sector, big data are being used to develop products that create value for firms and consumers. ZestFinance, using many more variables than traditional credit scoring, helps lenders determine whether or not to offer small, short-term loans to people who are otherwise poor credit risks.[31] This provides a better alternative to people who otherwise might rely on payday lenders or even loan sharks. LendUp, BillFloat, and ThinkFinance are companies following similar models that can provide better loan options for lower-income consumers, while Kabbage and On Deck Capital provide lending services to very small businesses.[32]

Two successful startups, Farecast, purchased by Microsoft, and Decide.com, recently purchased by eBay, use big data to help consumers find the lowest prices.[33] Farecast uses billions of flight-price records to predict the movement of airfares, saving purchasers an average of $50 per ticket. Decide.com predicts price movements for millions of products with potential savings for consumers of around $100 per item.

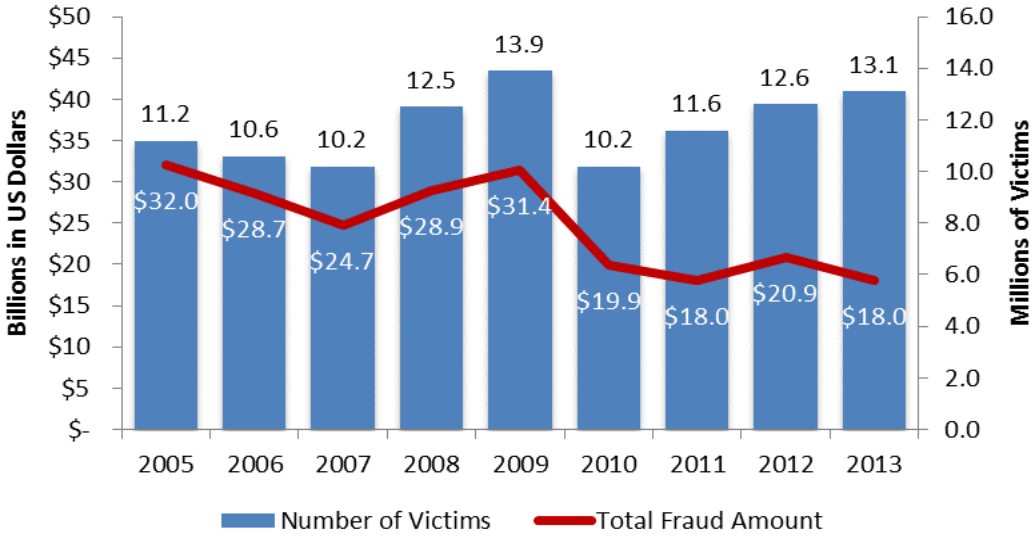
Another new company, Factual, collects data on over 65 million user locations and combines them with other data to help provide location-specific services, content and advertising.[34]

Big data are also used to protect against adverse events ranging from credit card fraud to terrorism. As Mayer-Schönberger and Cukier note, "the detection of credit card fraud works by looking for anomalies, and the best way to find them is to crunch all the data rather than a sample."[35] Einav and Levin cite a "Palo Alto company, Palantir, [which] has become a multi-billion dollar business by developing algorithms that can be used to identify terrorist threats using communications and other data, and to detect fraudulent behavior in health care and financial services."[36] They also cite work from a group at Dartmouth using large samples of Medicare claims to demonstrate substantial unexplained variation in Medicare spending per enrollee that could be due to inefficiencies or fraud.[37]

---

[29] Hyunyoung Choi and Hal Varian, Predicting the Present with Google Trends, Economic Record, Vol. 88, pp. 2-9.
[30] Lynn Wu and Erik Brynjolfsson, "The Future of Prediction: How Google Searches Foreshadow Housing Prices and Quantities", *ICIS 2009 Proceedings*, Paper 147, 2009, http://aisel.aisnet.org/icis2009/147.
[31] See an explanation of ZestFinance's techniques at http://www.zestfinance.com/how-we-do-it.html and a discussion of early success for the strategy at http://techcrunch.com/2013/07/31/data-focused-underwriting-and-credit-analysis-platform-zestfinance-raises-20m-from-peter-thiel-and-others/. Also, see Mayer-Schönberger and Cukier, p. 47.
[32] See company pages at https://www.lendup.com/; https://www.billfloat.com/; http://www.thinkfinance.com/; https://www.kabbage.com/; and https://www.ondeck.com/.
[33] Mayer-Schönberger and Cukier, p. 124.
[34] See discussions of Factual's growth and methods in interviews with its CEO at http://streetfightmag.com/2013/06/05/with-disparate-data-factual-founder-sees-opportunity/ and http://www.adexchanger.com/mobile/factual-eyes-new-opportunities-in-location-data/.
[35] Mayer-Schönberger and Cukier, p. 27.
[36] Einav and Levin, p. 7.
[37] Einav and Levin, pp. 10-11.

Figure 3: Overall Identity Fraud Incidence Rate and Total Fraud Amount by Year



Source: *2013 & 2014 Identity Fraud Report Sample*, Javelin Strategy and Research

Many of the innovations described above use multiple sources of data, which involves transferring data to third parties. Combining different data sets can greatly enhance their value for purposes ranging from epidemiology studies (e.g., the Danish study of cellphone use and cancer risk) to marketing. A recent study from the Direct Marketing Association found that individual-level consumer data were an integral component in producing over $150 billion in marketing services and that over 70 percent of these services required the ability to exchange data between firms.[38] These marketing services reduce the cost of matching producers with potential consumers in a marketplace, and are particularly valuable to smaller firms and new entrants.

III. Potential Privacy Threats

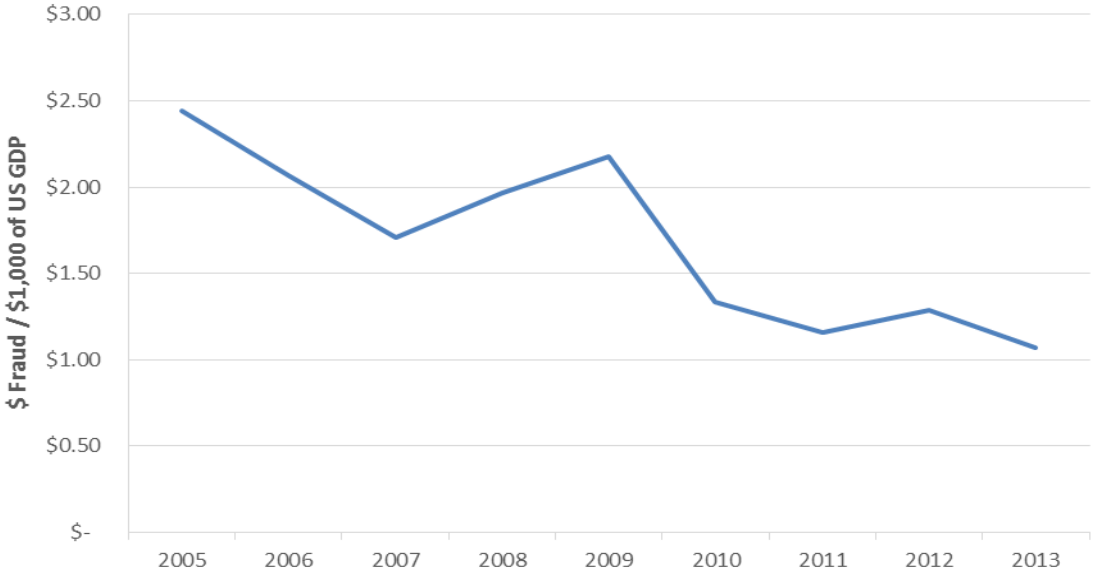
Privacy advocates, scholars and public officials have raised concern over a number of potential privacy threats from big data. As of now there is no evidence that any of these threats has materialized. We discuss them in turn.

Big data increase the risks associated with identity fraud and data breaches

In her speech referenced above, Chairwoman Ramirez suggests that big data increase the risks associated with identity fraud and data breaches.[39] These security issues might indicate a market

[38] John Deighton and Peter A. Johnson, "The Value of Data: Consequences for Insight, Innovation & Efficiency in the US Economy", *The Data Driven Marketing Institute*, October, 2013, available at http://ddminstitute.thedma.org/#valueofdata.
[39] Ramirez, p. 6.

Figure 4: Annual Cost of Identity Fraud (in dollars) Deflated by US GDP



Sources: *How Consumers Can Protect Against Identity Fraudsters in 2013,* Javelin Strategy and Research; Federal Reserve Economic Data, Federal Reserve Bank of St. Louis

failure because of the difficulty of imposing costs on the perpetrators, who may be able to remain anonymous or out of the reach of U.S. law enforcement.

In theory, big data could increase or decrease identity fraud and data breaches. On the one hand, there are simply more data at risk. On the other hand, the data themselves are useful in preventing fraud. Moreover, countervailing forces provide strong incentives for data holders (e.g., credit card companies) to protect their data. It is useful, therefore, to examine whether the proliferation of data in recent years has shown up in greater incidence of identity fraud and/or data breaches.

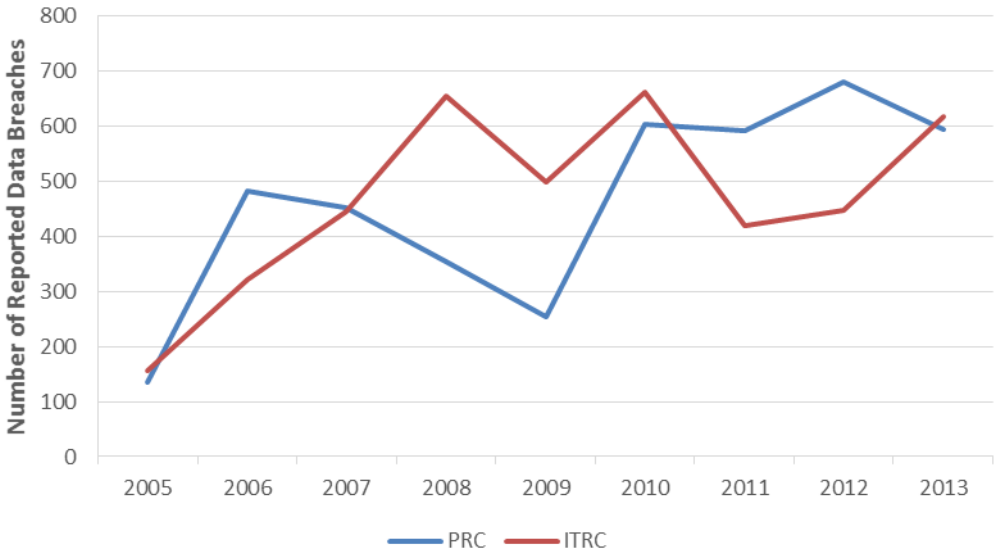
Identity Fraud

Javelin Strategy and Research compiles the only statistically representative series on identity fraud of which we are aware. These data are presented in Figure 3. Contrary to concerns voiced by the FTC and others, the overall incidence of identity fraud has been relatively flat since 2005. During the same period, the annual dollar amount of fraud has fallen—from an average of $29.1 billion for 2005-2009 to $19.2 billion for 2010-2013.

To obtain a clearer picture of what has happened to the risk of identity fraud, we need to normalize the data on identity fraud by some measure of exposure.[40] Figure 4 shows that the cost of identity fraud per $1,000 of US GDP has been declining since 2005. If the identity fraud

[40] This is the same thing analysts do when examining, for example, the risks associated with driving. They don't simply look at the number of accidents. They look at the number of accidents per mile driven.

Figure 5: Number of US Data Breaches per Year



Sources: *Data Breach Reports,* Identity Theft Resource Center; *Chronology of Data Breaches,* Privacy Rights Clearinghouse

cost data were deflated by e-commerce retail sales the downward trend would be steeper, because e-commerce has grown more rapidly than GDP. However, GDP is probably a more appropriate deflator, since the great majority of identity fraud is due to offline behavior.[41]

Data Breaches

There are two sources of data on data breaches—the Privacy Rights Clearinghouse and the Identity Theft Resource Center. Both of these sources collect aggregate information on data breaches from the media, public databases, and news releases from state governments; however, the annual totals vary slightly based on methodology and their individual definitions of a data breach.[42] Figure 5, which shows both series, suggests that the trend is slightly up since 2005. Data breaches are purely an online phenomenon, so it is appropriate to deflate them by a measure of online activity. When deflated by the volume of e-commerce, the risk of a data breach has been relatively constant, as shown in Figure 6.

Because data breaches can range from a handful to millions of stolen records, a more important measure is arguably the number of records compromised and the number of records compromised deflated by some measure of exposure such as e-commerce dollars. These are

[41] Only about 15 percent is associated with data breaches and online causes. The remainder is due to offline causes, including a stolen wallet or purse, auto burglary, home burglary and signature forgery. See Travelers Insurance, "73% of identity fraud cases resulted from stolen personal items", November 2012, Online, available at: http://investor.travelers.com/phoenix.zhtml?c=177842&p=irol-newsArticle&ID=1761670.

[42] For more information on the ITRC and PRC databases see http://www.idtheftcenter.org/id-theft/data-breaches.html and https://www.privacyrights.org/data-breach-FAQ.

Figure 6: Number of US Data Breaches per Year Deflated by US E-Commerce

Sources: *Data Breach Reports*, Identity Theft Resource Center, *Chronology of Data Breaches*, Privacy Rights Clearinghouse; *Quarterly E-Commerce Reports*, US Census Bureau

shown in Figures 7 and 8, respectively.[43] The spikes in records breached in 2007 and 2009 are due to three major breaches—TJ Maxx in 2007 (100 million records) and Heartland Payment Systems (130 million records) and a military veterans database (76 million records) in 2009. Overall, the trend in records breached since 2005 is relatively constant or even declining slightly, and the trend in records breached deflated by e-commerce volume is somewhat more negative.

Although the data on identity fraud and breaches are far from complete, there is no indication that either has gone up with the rise of big data. Indeed, one would expect that the use of big data might reduce identity fraud. This is because credit card companies, which bear most of the costs, have strong incentives to police misuse of their cards. One obvious method is monitoring purchases and notifying consumers when purchases seem to be outside of normal behavior, as determined by analysis of big data. Note that this policing involves use of data for purposes other than for which they were initially collected.

The use of big data to develop predictive models is harmful to consumers

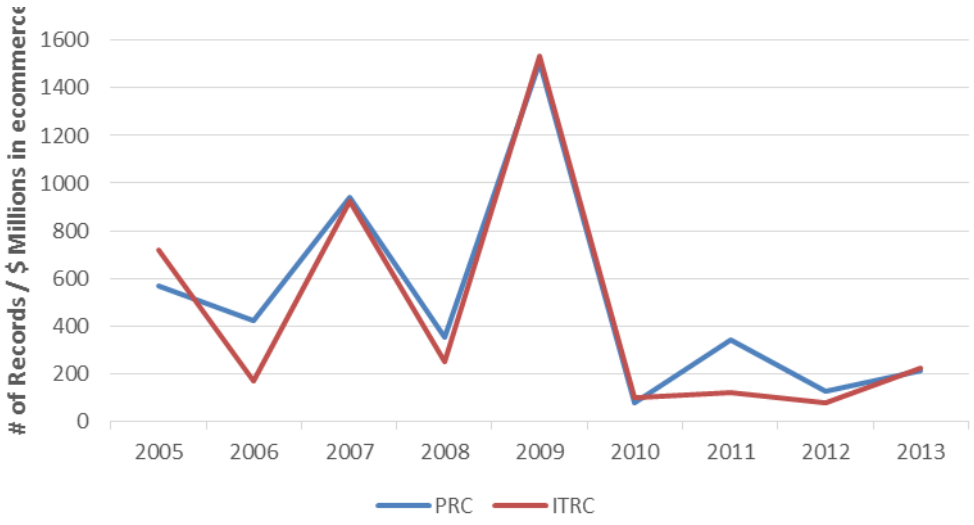
The systematic use of individuals' data for a wide range of purposes is not new. The direct marketing industry has for decades assembled mailing lists of consumers interested in specific products and services. Credit bureaus use formulas that determine individuals' eligibility for loans and the rates they may be offered. Similarly, the insurance industry uses key variables that indicate risk to determine whether and at what rates to offer insurance policies.

[43] Note these values should be viewed with some caution as the number of records compromised is not known for every reported breach. In fact, the percentage of reports with a known number of records has varied from 30% in some years to 87% in others.

Figure 7: Number of Reported Individual Records Compromised by Data Breaches

Sources: *Data Breach Reports*, Identity Theft Resource Center; *Chronology of Data Breaches*, Privacy Rights Clearinghouse

Figure 8: Number of Reported Individual Records Compromised by Data Breaches Deflated by US E-Commerce



Sources: *Data Breach Reports*, Identity Theft Resource Center, *Chronology of Data Breaches*, Privacy Rights Clearinghouse; *Quarterly E-Commerce Reports*, US Census Bureau

A theme permeating the privacy-centric big data literature is that the use of data to develop predictive models is harmful to consumers. As Chairwoman Ramirez said, "There is another risk that is a by-product of big data analytics, namely, that big data will be used to make determinations about individuals, not based on concrete facts, but on inferences or correlations that may be unwarranted."[44] She notes that "Individuals may be judged not because of what they've done, or what they will do in the future, but because inferences or correlations drawn by algorithms suggest they may behave in ways that make them poor credit or insurance risks, unsuitable candidates for employment or admission to schools or other institutions, or unlikely to carry out certain functions." She further points out, "An error rate of one-in-ten, or one-in-a-hundred, may be tolerable to the company. To the consumer who has been miscategorized, however, that categorization may feel like arbitrariness-by-algorithm."[45]

This point has also been made by Commissioner Brill: "They [data brokers] load all this data into sophisticated algorithms that spew out alarmingly personal predictions about our health, financial status, interests, sexual orientation, religious beliefs, politics and habits…. [I]ncreasingly our data fuel more than just what ads we are served. They may also determine what offers we receive, what rates we pay, even what jobs we get."[46]

Such criticism, however, applies to quantitative analysis used for decision-making throughout the economy. Use of credentials and test scores is universal in American life. For example, the educational testing industry is based on the use of such correlations.[47] The Federal Government, including the FTC, uses class rank in hiring lawyers. These decisions are based on "small data"—sometimes, one test score or one data point. Big data can only improve this process. If more data points are used in making decisions, then it is less likely that any single data point will be determinative, and more likely that a correct decision will be reached.

It is important to emphasize that companies devote resources to gathering data and undertaking complex analysis because it is in their interest to make more accurate decisions. Sometimes that involves discovering that seemingly unrelated variables are, in fact, related. Thus, big data should lead to fewer consumers being mis-categorized and less arbitrariness in decision-making.

It is unclear what kinds of "inferences or correlations may be unwarranted." Insurance companies typically give a discount on auto insurance to students with good grades, for example. They also differentiate on the basis of the gender of young drivers. This is presumably because the data show a correlation between these variables—school performance and gender—and accident costs.[48]

The use of more variables made possible by big data should lead to more accurate decisions that also might be "fairer." For example, ZestFinance, described above, uses its big data analysis to

---

[44] Ramirez, p. 7.
[45] Ramirez, p. 8.
[46] Brill, ¶ 3,4.
[47] See, for example, William A. Mehrens, "Using Test Scores for Decision Making", *Test Policy and Test Performance: Education, Language, and Culture*, 1989, pp. 93-113.
[48] See for example: http://www.cnbc.com/id/100863117.

help underwrite loans to individuals who would otherwise not qualify. Another example is the greater use of data by state parole boards to help inform parole decisions.[49] Proponents believe the use of big data in this manner provides more accurate predictions of the risk of recidivism and therefore can help determine which prisoners should be released and thereby increase public safety and perhaps also reduce prison costs.

The use of big data is discriminatory

Some writers argue that the use of big data in marketing decisions favors the rich over the poor.[50] A few particularly inflammatory quotes from critics include: "Ever-increasing data collection and analysis have the potential to exacerbate class disparities;"[51] and, "Big data—discrimination, profiling, tracking, exclusion—threaten the self-determination and personal autonomy of the poor more than any other class."[52] One writer theorized, "To woo the high value shoppers, they offer attractive discounts and promotions—use your loyalty card to buy Beluga caviar; get a free bottle of Champagne. Yet obviously the retailers can't take a loss for their marketing efforts. Who then pays the price of the rich shoppers' luxury goods? You guessed it, the rest of us—with price hikes on products like bread and butter."[53]

The argument that data collection favors the rich over the poor is presented without evidence. The example of consumption of caviar and Champagne by rich people being subsidized by price increases on bread and butter is, as far as we can tell, hypothetical.

Likely the concern expressed by these writers relates to price discrimination, which involves charging different prices to different consumers for the same product based on their willingness to pay.[54] Online data collection can facilitate price discrimination, because it yields information that can be used to infer a consumer's willingness to pay for a good.[55]

Price discrimination transfers some (or even all in the case of perfect price discrimination) surplus from consumers to producers. However, price discrimination is economically efficient (i.e., increases welfare overall) if it increases total output in a market. Particularly in the case of products with high fixed and low marginal costs—such as airline tickets—price discrimination may be necessary for the good to be produced at all. There would be fewer flights if airlines were not able to charge varying prices. Many virtual goods, such as apps and software, also have

---

[49] See Joseph Walker, "State Patrol Boards Use Software to Decide Which Inmates to Release", *The Wall Street Journal Online*, October 12, 2013, available at: http://online.wsj.com/news/article_email/SB10001424052702304626104579121251595240852-lMyQjAxMTAzMDEwMDExNDAyWj.

[50] This concern is reflected in a forthcoming FTC workshop on the effects of big data on low income and underserved consumers. See announcement at http://www.ftc.gov/news-events/press-releases/2014/04/ftc-examine-effects-big-data-low-income-underserved-consumers?Source=govdelivery.

[51] Joseph Jerome, "Buying and Selling Privacy: Big Data's Different Burdens and Benefits", 66 Stanford Law Review Online 47, 2013, p. 50.

[52] Jerome, p. 51.

[53] Omer Tene, "Privacy: For the Rich or for the Poor", Concurring Opinions Blog Post, June, 2012, available at: http://www.concurringopinions.com/archives/2012/07/privacy-for-the-rich-or-for-the-poor.html.

[54] See Tene, ¶ 4,6.

[55] Hal R. Varian, "Differential Pricing and Efficiency", *First Monday* Vol. 1, No. 5, August, 1996, http://www.firstmonday.dk/ojs/index.php/fm/article/view/473/394.

high fixed and low or even zero marginal costs, and price discrimination may be essential to the production of these goods.

Price discrimination involves charging prices based on a consumer's willingness to pay, which in general is positively related to a consumer's ability to pay. This implies that a price discriminating firm will, other things the same, charge lower prices to lower-income consumers. Indeed, in the absence of price discrimination, some lower-income consumers would be unable or unwilling to purchase some products at all. So, contrary to arguments above, the use of big data, to the extent it facilitates price discrimination, should usually work to the advantage of lower-income consumers.

Perhaps the most publicized conclusion of the recently released EOP report concerns the possibility of discrimination against vulnerable groups—that "big data analytics *have the potential to* (italics added) eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace."[56] However, the two examples of discrimination cited turn out to be almost non-examples.

The first example involves StreetBump, a mobile application developed to collect information about potholes and other road conditions in Boston. Even before its launch, the city recognized that this app, by itself, would be biased toward identifying problems in wealthier neighborhoods, because wealthier individuals would be more likely to own smartphones and make use of the app. As a result, the city adjusted accordingly to assure reporting of road conditions was accurate and consistent throughout the city.[57]

The second example involves the E-verify program used by employers to check the eligibility of employees to work legally in the United States. The report cites a study that "found the rate at which U.S. citizen (sic) have their authorization to work be initially erroneously unconfirmed by the system was 0.3 percent, compared to 2.1 percent for non-citizens. However, after a few days many of these workers' status was confirmed."[58] It seems almost inevitable that the error rate for citizens would be lower, because citizens automatically are eligible to work, whereas additional information is needed to confirm eligibility for non-citizens (i.e., evidence of some sort of work permit). Hence, it is not clear this is an example of discrimination.

Firms use big data to manipulate consumers

Some recent privacy literature suggests that the use of data and algorithms may produce "harms" quite different from what we normally think of as privacy and security harms (i.e., harms that involve the exposure of individuals' data to people who shouldn't see them). In a recent article,

[56] Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values,* May, 2014, Cover Letter.
[57] EOP Report, pp. 51-52.
[58] EOP Report, p. 52.

Calo hypothesizes that big data will help firms discover opportunities to capitalize on irrational behavior.[59]

Calo is concerned with "the 'mass production of bias' through big data,"[60] when "firms start looking for vulnerability."[61] Big data would be used as follows: "The first step is to model what rational choice would look like in a given context: consumers taking every realistic opportunity to maximize their own welfare. The second step is to analyze consumer transactions by the millions to spot the places where consumers have deviated from the rational model in small and big ways."[62] Calo acknowledges that "modeling 'rational' behavior would be difficult,"[63] but ultimately he believes that "A firm with the resources and inclination will be in a position to surface and exploit how consumers tend to deviate from rational decision making on a previously unimaginable scale. Firms will increasingly be in the position to *create* suckers, rather than waiting for one to be born every minute."[64] He poses the question: "When does personalization… become an issue of consumer protection?"[65]

Drawing a boundary between what is called "manipulation" and the provision of information that helps a consumer in making purchases is difficult, as Calo acknowledges: "Obviously manipulating consumers is not the only—nor, for many, the primary—use to which firms will put consumer data. Data helps firms improve existing products and develop the indispensable services of tomorrow. Data is necessary to combat various kinds of fraud and sometimes to police against one set of consumers abusing another. Regulators are rightfully concerned about the effects of cutting off data flows on innovation. Telling services what data they can and cannot collect, meanwhile, creates pragmatic line-drawing problems that regulators may not be well-suited to answer."[66]

Calo suggests "regulators and courts should only intervene where it is clear that the incentives of firms and consumers are not aligned."[67] As an example, a harmful use of information would be to send an obese consumer a text message from a donut shop when the consumer is trying to avoid snacking.[68] But of course the consumer might want a donut, even though Calo thinks he should not have one. Moreover, given the rate of evolution of apps, there will soon be one (if there is not now) that a diet conscious consumer with weak willpower could program to ignore all messages with certain keywords, including "donut", or to remind him of the caloric content of the donut and his current weight-loss goal.

---

[59] Ryan Calo, "Digital Market Manipulation", Legal Studies Research Paper and George Washington Law Review (forthcoming), 2013, University of Washington School of Law, available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2309703.
[60] Calo, DMM, p. 12
[61] Calo, DMM, p. 15
[62] Calo, DMM, p. 15
[63] Calo, DMM, p. 15
[64] Calo, DMM, p. 22
[65] Calo, DMM, p. 5
[66] Calo, DMM, pp. 44-45.
[67] Calo, DMM, p. 26.
[68] Calo, DMM, p. 3.

As Calo also acknowledges, profiting from irrational behavior would be difficult (perhaps impossible) since it would be extremely difficult to determine what is rational for a given consumer.[69] Moreover, there is no clear reason why firms would want to do this. Using large data sets, firms might simply determine when they can sell products, and most of the time that would be to consumers who want the product, and that would generally be to rational consumers. Moreover, while some firms might try to sell products that the consumer does not "really" want, others would be trying to sell products that the consumer does want, and those firms can be expected to win out. An implicit assumption in Calo's discussion is a lack of competition. Even assuming firms can manipulate consumers and thereby earn super-competitive profits, unless there are barriers to entry, other firms will be induced to enter and compete away those super-competitive profits. This is a check on whatever manipulation might be possible.

In general, it is not possible to determine whether any given purchase is "rational" or not, because consumers' utility functions are not directly observable. In a market economy, firms are rewarded for giving consumers what they want. The economist's criterion of performance is how close the economy comes to maximizing "total surplus."

It is true that firms want to capture as much of that surplus as possible, and in that sense, their interests may not be aligned with those of consumers. Calo is concerned that firms will use data to find that moment of vulnerability when they can charge consumers a higher price. Two observations on this: First, the transaction will still be beneficial to the consumer; she may just capture less consumer surplus. Second, this is also a way that firms can efficiently price discriminate (see discussion above). Others may get a lower price. Importantly, such price discrimination may be necessary to cover costs and for the product to be available at all. In a competitive market, price discrimination that leads to excess profits will attract entry.

Big data will reduce the variety of information consumers see

Some writers express concern about consumers living in a big-data-facilitated "filter bubble" because of predetermined interests. As a result, consumers would not be exposed to a wide variety of information or services they may find useful.

For example, Pariser laments that "The statistical models that make up the filter bubble write off the outliers. But in human life it's the outliers who make things interesting and give us inspiration."[70] Dwork and Mulligan are concerned that "filter bubbles" will take away "the tumult of traditional public forums—sidewalks, public parks, and street corners—where a measure of randomness and unpredictability yields a mix of discoveries and encounters that contribute to a more informed populace."[71]

---

[69] Calo, DMM, p. 15.
[70] Eli Pariser, "The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think", Penguin Books, April, 2012, p. 134.
[71] Cynthia Dwork and Deirdre K. Mulligan, "It's Not Privacy and It's Not Fair", 66 Stanford Law Review Online 47, 2013, p. 39.

If consumers want variety, big data and algorithms, particularly as they get more sophisticated, should be helpful in providing that variety to them. However, the notion that algorithms will give consumers "too much" of what they want at the expense of what is good for them is a more radical idea with unclear policy implications. Does it mean we should limit the collection and use of data to purposely produce less accurate algorithms? The fundamental problem with this line of analysis is that many of the privacy advocates and writers on this subject do not seem to trust the judgment of consumers for whom they purport to advocate to make choices.

Individuals will be forced to reveal data about themselves, thereby eroding privacy

As described above, the systematic gathering and use of data (by algorithm as well as less formal means) to determine eligibility for credit, employment and insurance as well as for marketing purposes is ubiquitous. The flip side of this "sorting" is "signaling," in which individuals (or firms) voluntarily reveal information about themselves in order to address asymmetric information problems. As Michael Spence notes, "The incentive to engage in activities that inform buyers is greatest for sellers with high quality products, but if they are successful, the incentive will trickle down through the spectrum of qualities."[72]

In a recent paper, Scott Peppet suggests that "the Internet and digitization are decreasing the transaction costs of signaling by making verifiable signals more readily available throughout the economy, and signaling will thus continue to become more and more important and ubiquitous as a response to information asymmetries."[73] With advanced information technologies, individuals will increasingly be able to voluntarily make available a range of *verified* information (i.e., linked directly to the source) about themselves, including health information, employment records, court records, driving behavior and credit history. For example, your health data may be generated by wearable monitors, your driving behavior by sensors in your car, etc.

The data made available could determine eligibility and the terms for many economically important items, including jobs, insurance and admission to schools. Those with the most favorable data will find it in their interest to reveal it. Others will then be "forced" to reveal their data because failure to do so will reflect negatively on those who do not. This, according to Peppet, "contains within it a radical threat: the possible unraveling of privacy altogether as some individuals initially find it in their interest to disclose information for personal gain and then, as the unraveling proceeds, all realize that disclosure is no longer a choice but instead a necessity as the signaling economy attaches a stigma to staying silent."[74] Moreover, "rapidly changing information technologies are making it possible for consumers to share verified personal information at low cost for economic reward or, put differently, for firms to extract previously unavailable personal information from individuals… [which] poses a very different threat to

[72] Michael Spence, *Informational Aspects of Market Structure: An Introduction,* The Quarterly Journal of Economics, Vol. 90, No. 4, November 1976, p. 592.
[73] Scott Peppet, *Unraveling Privacy: The Personal Prospectus and the Threat of a Full-Disclosure Future,* Northwestern University Law Review, Vol. 105, No. 3, p. 1166.
[74] Peppet, p. 1176

privacy than the threat of data mining, aggregation and sorting that has preoccupied the burgeoning information privacy field for the last decade."[75]

As indicated by the Ackerloff and Spence articles, this "unraveling" phenomenon has been the subject of study for quite a while.[76] Generally, it is considered efficient, precisely because it provides information to the market. In the absence of information, markets may "unravel" in another way. "An important implication of [Ackerloff's lemons paper] was that high quality sellers may withdraw their products from the market because their products cannot be distinguished and therefore are priced according to the average."[77]

Posner has written in opposition to mandates that protect certain types of information, arguing there is a "symmetry between 'selling' oneself and selling a product. If fraud is bad in the latter context—at least to the extent that we would not think it efficient to allow sellers to invoke the law's assistance in concealing defects in their goods—it is bad in the former context, and for the same reasons: it reduces the amount of information in the market, and hence the efficiency with which the market… allocates resources."[78] Moreover, "Once privacy is seen to reduce the efficiency of the marketplace, we are in a position to predict the effect of the recent wave of statutes, federal and state, protecting privacy, as by placing arrest records beyond a prospective employer's reach and credit histories beyond a prospective creditor's reach. If the analysis in this paper is correct, such statutes reduce wages and employment and increase interest rates."[79]

In the same way that prohibiting producers from hiding defects in their products leads to better products, there are positive incentive effects when individuals are unable to conceal adverse personal information. The fact that a better grade point average lowers automobile insurance rates for young males is an incentive to study harder, or, at least, for parents to make sure their student studies harder. If individuals were able to conceal their credit histories, we would find more delinquent payments, which would raise the costs of borrowing generally. The fact that having a criminal record makes it difficult to find a job is likely some deterrent to criminal behavior.

A simple example illustrates the potential costs of restricting this type of information sharing. It is now possible to monitor driving behavior for a variety of purposes. Mapping programs do this in order to direct drivers to the fastest route at any given time. A company called Automatic helps people monitor their driving in order to reduce costs by improving fuel economy and reducing wear and tear.[80]

---

[75] Peppet, pp. 1155-1156.
[76] Ackerloff and Spence, along with Joseph Stiglitz, received the 2001 economics Nobel Prize for their study of markets with asymmetric information.
[77] Spence, p. 591.
[78] Richard A. Posner, "The Economics of Privacy," Center for the Study of the Economy and the State, University of Chicago, Working Paper No. 016, October 1980, p. 5.
[79] Posner, p. 5.
[80] See, http://www.automatic.com/.

Similar devices can also be used by insurance companies to set rates. A driver can install a monitor in her car and have the data automatically delivered to her insurance company.[81] Presumably, safe drivers will want to do this so they can get lower rates. Insurance companies might rationally respond by assuming that drivers who failed to install such devices were less safe than those who did and charging them a higher rate. This would likely result in at least a partial "unraveling" as more and more drivers installed the monitoring devices.

Prohibiting this practice, as some privacy advocates suggest, would mean there is no payoff to voluntarily providing your monitoring data to the insurance company. This would penalize safe drivers to the benefit of average and less-safe drivers. The prohibition would increase accidents, because even the safest drivers will drive more carefully when they know they are being monitored. There may be a significant increase in accidents from drivers further down the spectrum, who otherwise would be induced to install a monitor.

Finally (and somewhat ironically), Peppet does not discuss what might seem to be the most obvious application of signaling in the context of privacy—the ability of firms to compete by offering better privacy policies. Privacy is a quality attribute, just like any other. If consumers value it, those firms with the "best" privacy practices might be expected to advertise that fact, forcing a general "unraveling" of privacy practices that would inform consumers. The fact that firms compete less on the basis of privacy than we might expect suggests that consumers are less concerned about privacy practices than other firm attributes.

IV. Policy Considerations

There is no obvious reason to approach privacy policy questions arising from big data differently than we approach questions involving smaller amounts of data. The same questions are relevant.

First, policy makers should ask whether there is a market failure or evidence of harm to consumers. The recent literature on big data we have surveyed does not provide such evidence, at least as far as the legal use of data for commercial purposes is concerned.[82] Discussions of harm in the literature are largely speculative and hypothetical. Moreover, we have found no evidence of an increase in harm to consumers from identity fraud or data breaches.

Some examples of what have been described as "objective privacy harms" include: use of blood test data for drunk driving; data used for a no-fly list; and police use of information from a psychologist.[83] Only some of these are related to big data, but more importantly, none involve commercial information. They all involve government functions, which most people would think are legitimate. The only example citing commercial use is from Google gmail ads. But in

---

[81] See, http://www.progressive.com/auto/snapshot-how-it-works/.

[82] This is consistent with our conclusion that demonstrable harm from the legal use of commercial information is lacking. See Thomas Lenard and Paul Rubin, "In Defense of Data: Information and the Costs of Privacy", *Policy & Internet*, Vol. 2, Issue. 1, April, 2010, pp. 149-183, available at http://onlinelibrary.wiley.com/doi/10.2202/1944-2866.1035/abstract.

[83] Ryan Calo, "The Boundaries of Privacy Harm", *Indiana Law Journal,* Vol. 86, No. 3, 2011, pp. 1131-1162.

this case, the consumer voluntarily uses the service in full knowledge that he will receive targeted ads in exchange for a free product. Moreover, the "harm" identified is speculative and quite indirect—consumers using the service are not typically aware of any harm.

If evidence of market failure or harm is found, the next question for policy makers is whether an available remedy (or remedies) can reasonably be expected to yield benefits greater than costs and therefore net benefits to consumers. This, in turn, leads to the threshold question of whether there are harms that can be reduced by the implementation of a new privacy policy. Otherwise, there can be no benefits. Since the privacy harms cited in the literature are largely hypothetical, so are the benefits. In other words, the absence of identified harms implies that privacy policies cannot be expected to yield net benefits, even in the absence of costs.

The privacy remedies typically discussed are, however, likely to impose costs. A standard solution long promoted by privacy advocates is that data should only be collected for a specific, identified purpose. This is reflected in the FIPPs dating back to the 1970s, the OECD Privacy Principles of 1980, current European Union regulations, and the recommendations of the FTC's 2012 Privacy Report.[84] Indeed, according to Chairwoman Ramirez, the First Commandment of data hygiene is: "Thou shall not collect and hold onto personal information unnecessary to an identified purpose."[85] Similarly, Commissioner Julie Brill laments the fact that firms, "without our knowledge or consent, can amass large amounts of private information about people to use for purposes we don't expect or understand."[86]

Chairwoman Ramirez's First Commandment is particularly ill-suited to the world of big data and, in fact, is inconsistent with other parts of her speech where she points out beneficial uses of big data, such as: improving the quality of health care while cutting costs, making more precise weather forecasts, predicting peak electricity consumption, and delivering better products and services to consumers at lower costs.[87] These beneficial uses often involve using medical data, utility billing records and other data for purposes other than those for which they were initially collected.

Moreover, the government itself routinely violates the data-hygiene First Commandment. When people paid their taxes, for example, they did not know that data from their returns would later be used to determine their eligibility for health insurance subsidies. Indeed, individuals could not have been informed of that potential use when they filed their returns, as using the data in such a way was only recently envisioned.

Using data in unanticipated ways has been a hallmark of the big data revolution, for commercial, research and even public sector uses. Therefore, policies that limit the reuse or sharing of data

---

[84] An excellent summary of the evolution of the FIPPs comes from Robert Gellman, "FAIR INFORMATION PRACTICES: A Basic History", last updated November 11, 2013, available at http://www.bobgellman.com/rg-docs/rg-FIPShistory.pdf, and the current FTC FIPPs are posted at http://www.ftc.gov/reports/privacy3/fairinfo.shtm.
[85] Ramirez, p. 4.
[86] Julie Brill, "Demanding transparency from data brokers", *Washington Post Opinions*, August 15, 2013.
[87] Ramirez, p.1.

would be particularly harmful if applied to big data because they are inconsistent with the innovative ways in which data are used.

Principles of notice and choice become almost meaningless when data may be used in unpredictable ways. Even absent questions concerning big data, these principles have become increasingly irrelevant. As Beales and Muris note, "The reality that decisions about information sharing are not worth thinking about for the vast majority of consumers contradicts the fundamental premise of the notice approach to privacy." They continue, "The FIPs principle of choice fares no better." [88]

Both of the recently released White House reports indicate that the FIPPs focus on limiting data collection is increasingly irrelevant and, indeed, harmful in a big data world. The EOP report observes that "these trends may require us to look closely at the notice and consent framework that has been a central pillar of how privacy practices have been organized for more than four decades."[89] The PCAST report notes, "The beneficial uses of near-ubiquitous data collection are large, and they fuel an increasingly important set of economic activities. Taken together, these considerations suggest that a policy focus on limiting data collection will not be a broadly applicable or scalable strategy—nor one likely to achieve the right balance between beneficial results and unintended negative consequences (such as inhibiting economic growth)."[90]

Transparency

Concern about the use of data for predictive scoring and the possibility that algorithms may mis-categorize individuals sometimes leads to recommendations for greater transparency and "procedures to remediate decisions that adversely affect individuals who have been wrongly categorized by correlation."[91] This is the thrust of Commissioner Brill's "Reclaim Your Name" initiative. One major data broker, Acxiom, has taken a step in that direction with its aboutthedata.com web site, which allows individuals to view and potentially correct some of the data in Acxiom's file.[92]

The notion that consumers should understand who is collecting their data and how they are being used is an appealing one, but it is largely meaningless, especially in the big data era where scores may be based on hundreds of data points and very complex calculations. For example, it is not clear that a person rejected for credit by a complex algorithm would particularly benefit by being shown the equation used. The FICO score, an early example of a calculation based on a complex

---

[88] J. Howard Beales and Timothy Muris, *Choice or Consequences: Protecting Consumer Privacy in Commercial Information*, University of Chicago Law Review, 2008, pp. 113-118.
[89] EOP Report, p. 54.
[90] PCAST Report, pp. x-xi.
[91] Ramirez, p. 8.
[92] This effort, however, has been criticized by privacy advocates as being too limited. See Natasha Singer, "Acxiom Lets Consumers See Data It Collects", *The New York Times*, September 4, 2012, available at: http://www.nytimes.com/2013/09/05/technology/acxiom-lets-consumers-see-data-it-collects.html.

algorithm, is virtually impossible to explain to even an informed consumer because of interactions and nonlinearities in the way various data points enter into the score.[93]

Electronic information is frequently used in complex ways that are difficult or impossible to explain. It would not be feasible for websites to meaningfully convey this information through a notice, and consumers would not devote the hours required to understand it. For example, a Wall Street Journal series titled "What They Know" consisted of several lengthy articles explaining uses of data. Indeed, from the articles, it appears that many practitioners do not themselves understand the ways in which they are using data.[94] Rubin and Lenard present a complicated schematic showing the uses of data as of 2001.[95] Since then uses of data have become even more complex.

Giving consumers the ability to correct their information may be more complicated than it might appear, even aside from the administrative complexities. Consumers do have the right to correct information used in deriving their credit scores, but it is made difficult to do so, for good reason. An individual who thinks she has been wrongly categorized clearly has an interest in correcting erroneous information if that information has a negative effect. But she might also have an interest in "correcting" valid information that would adversely affect the decision, or inserting incorrect information that would have a positive effect. Distinguishing between these various "corrections" may be quite difficult.

The purpose of collecting information that affects decisions about individuals—e.g., credit decisions, insurance decisions, or employment decisions—is to ameliorate an asymmetric information problem. Individuals have much more information about themselves than lenders, insurance companies or prospective employers. As discussed in Section III, asymmetric information is a feature of some markets that potentially can lead to market breakdown.[96] This is why, as Beales and Muris point out, "In our economy, there are vital uses of information sharing [such as credit reporting] that depend on the fact that consumers cannot choose whether to participate."[97]

Moreover, if we make it easier for individuals to access their data then we also make it easier for those bent on fraud to access the same data. If fraudsters have access to large amounts of data about a person, they can more easily defraud that individual (perhaps by making purchases that are consistent with the individual's behavior in order to trick the credit card companies' monitoring efforts). Thus, ease of consumer monitoring is at best a two-edged sword.

---

[93] The major inputs to a credit score are well known; however, the calculation of credit scores from credit report data is proprietary and exceedingly complex. See for example FDIC, *Credit Card Activities Manual*, Ch. 8 – Scoring and Modeling, 2007, available at: http://www.fdic.gov/regulations/examinations/credit_card/.

[94] Julia Angwin, "The Web's New Gold Mine: Your Secrets", *The Wall Street Journal*, July 30, 2010, Online, available at: http://online.wsj.com/news/articles/SB10001424052748703940904575395073512989404.

[95] See Paul Rubin and Thomas Lenard, "Privacy and the Commercial Use of Personal Information", Kluwer Academic Publishers and Progress and Freedom Foundation, 2001, p. 26.

[96] See George A. Akerlof, "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism", *The Quarterly Journal of Economics*, Vol. 84, No. 3, August, 1970, pp. 488-500.

[97] Beales and Muris, p. 115.

Alternative Privacy Approaches

As an alternative to the standard FIPPs approach, Beales and Muris recommend an approach based on "the consequences of information use and misuse. There is little basis for concern among most consumers or policymakers about information sharing per se. There is legitimate concern, however, that some recipients of the information will use it to create adverse consequences for the consumer."[98] As an example of a consequence-based policy, they point to the Do Not Call Registry (aimed at the adverse consequence of receiving unwanted marketing calls) established when Muris was FTC Chairman and Beales was Director of the agency's Bureau of Consumer Protection.

The recently released White House reports reflect a similar approach. The EOP report suggests examining "whether a greater focus on how data is used and reused would be a more productive basis for managing privacy rights in a big data environment."[99] The PCAST report is even clearer:

> Policy attention should focus more on the actual uses of big data and less on its collection and analysis. By actual uses, we mean the specific events where something happens that can cause an adverse consequence or harm to an individual or class of individuals…. By contrast, PCAST judges that policies focused on the regulation of data collection, storage, retention, a priori limitations on applications, and analysis… are unlikely to yield effective strategies for improving privacy. Such policies would be unlikely to be scalable over time, or to be enforceable by other than severe and economically damaging measures.[100]

Calo has offered two new policy ideas. The first is a "thought experiment" based on the example of academic institutional review boards (IRBs).[101] Researchers proposing experiments involving human subjects are required to submit their project to an IRB to ensure that human subjects are appropriately protected. The IRBs use principles articulated in the "Belmont Report" published by a government taskforce. Calo suggests something similar—a Consumer Subject Review Board—for commercial uses of data, to assure that the subjects (i.e., consumers) are adequately protected. Standards for the review board would be developed by the FTC, the Department of Commerce or by industry itself.

In general, firms, certainly firms concerned about their reputations, can be expected to take into account consumer reactions to their data practices—whether formally through internal committees or less formally—consistent with their fiduciary obligations to shareholders. If Calo is suggesting something more regulatory, it should be subjected to a benefit-cost analysis, which has not been performed. Since the evidence of harm from the use of data is thin to non-existent, it is doubtful that any such regulation could pass a benefit-cost test.

---

[98] Beales and Muris, p. 118
[99] EOP Report, p. 61.
[100] PCAST Report, p. xiii.
[101] Ryan Calo, *Consumer Subject Review Boards: A Thought Experiment*, Stanford Law Review Online, Vol. 66, p. 97.

Calo also proposes a "paid option regime:"[102] "Imagine if major platforms such as Facebook and Google were obligated, as a matter of law or best practice, to offer a paid version of their service. For, say, ten dollars a month or five cents a visit, users could opt out of the entire marketing ecosystem."

This is an alternative that, of course, is available in the market, but perhaps not as often as critics of advertising, such as Calo, would like. There are probably several reasons for this. The "paid option" involves substantial transactions costs on both sides. In addition, it is likely that the value of the data consumers provide in the "free regime" is greater than what consumers would be willing to pay. If Calo's paid option were a regulatory requirement, the obvious question would be "at what price?" This might imply price regulation, which, especially given the huge variety of online services, would not be feasible. This is also not a proposal that could pass a benefit-cost test.

V. Conclusion

The basic idea behind the standard privacy remedies reflected in PBD, the FIPPs, and the OECD principles, that have been the focus of privacy policy discussions for several decades, is to limit the collection and use of information. These principles have become increasingly irrelevant as they have become increasingly familiar, even aside from the fact that they fail to address identifiable harms. Using data in unanticipated ways has been a hallmark of the big data revolution. The standard solutions that would limit the reuse or sharing of data would be particularly harmful if applied to big data because they are inconsistent with the innovative ways in which data are being used. This would have a detrimental impact on innovation in a variety of sectors, from marketing to credit markets to health research.

Regulators, such as Chairwoman Ramirez suggest "meaningful oversight" as a remedy for perceived harms to consumers, but we should note that the FTC has sometimes shown itself to be an overprotective steward, and has often reduced consumer welfare by excessive regulation of information.[103] Neither the FTC nor any other regulator has performed cost-benefit analysis on the FIPPs or any of its variations.[104] Given this lack of data and analysis, particularly in a new market such as the electronic use of information, it is much more likely that an uninformed regulator will stifle innovation rather than provide net benefits.

---

[102] Calo, DMM, p. 48

[103] Paul H. Rubin, "Regulation of Information and Advertising," Competition Policy International, Spring 2008, v. 4, No. 1, pp. 169-192.

[104] See Exec. Order 13563, Improving Regulation and Regulatory Review (Jan. 17, 2011), available at http://www.gpo.gov/fdsys/pkg/FR-2011-01-21/pdf/2011-1385.pdf. Also, see Thomas M. Lenard and Paul H. Rubin, "The FTC and Privacy: We Don't Need No Stinking Data," Antitrust Source, October, 2012, available at http://www.americanbar.org/content/dam/aba/publishing/antitrust_source/oct12_lenard_10_22 f.authcheckdam.pdf.