United States of America

Federal Trade Commission

The Power of Data

Remarks of Maureen K. Ohlhausen[1]
Commissioner, Federal Trade Commission

Georgetown University McCourt School of Public Policy and Georgetown Law Center
Privacy Principles in the Era of Massive Data
Washington, DC

April 22, 2014

I. Introduction

Thanks for inviting me to kick off what I am certain will be an insightful discussion of big data and its implications for privacy. As society has integrated and adopted increasingly powerful computers and pervasive communications networks, we have created massive amounts of data. This trend will continue as we move into the era of the Internet of Things, a universe of far-flung devices that will massively increase the amount of passive data collection. The tools that will enable us to collect and analyze this "big data" promise significant benefits for consumers, businesses, and government. I may be preaching to the choir a little here, as the mission of the McCourt School's Massive Data Institute is, and I quote, "to use 'big data' sets to increase understanding of society and human behavior and thus improve public policy decision-

[1] The views expressed in this speech are solely those of Commissioner Ohlhausen and are not intended to reflect the views of the Commission or any other Commissioner. I would like to thank Neil Chilson for his assistance in preparing this speech.

making."[2] Ultimately, I share the optimism of that mission statement. Although some potential uses of big data raise concerns about privacy and other values, we can address these concerns together, through a coalition of academics, regulators, businesses, and consumers. Big data is a tool; like all tools it has strengths and weaknesses. Keeping those strengths and weaknesses in perspective is important as we work together to adapt our laws, guidelines, best practices, customs, and society to integrate this new technology. As we adapt to big data, the FTC will serve an important role in protecting consumers and promoting innovation.

II. Keeping Big Data in Perspective

If everything you knew about big data came from news reports, you could be forgiven for thinking that big data is either a miracle cure for all of our most intractable social problems or a plague upon consumers. Some have called big data a "revolution in how we live, work, and think,"[3] or health care's big savior.[4] Others have labeled big data "the death of Internet privacy"[5] and many have compared it to the Tom Cruise movie "Minority Report."[6]

Of course, the reality is more complicated than the headlines. So what is big data? Some criticize the term as a meaningless buzzword, saying we should be specific: are we talking about "information transformation, storage, and retrieval? Machine learning and data mining? Pattern

---

[2] Letter from Edward Montgomery, Dean, McCourt School of Public Policy, to Maureen K. Ohlhausen, Commissioner, FTC, April 7, 2014.

[3] *See generally* VIKTO MAYER-SCHÖNBERGER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK (Mar. 2013).

[4] *See* Ricky Ribeiro, *Will Big Data Become the Big Savior of Health?,* http://www.biztechmagazine.com/article/2013/06/will-big-data-become-big-savior-health (last visited Apr. 21, 2014).

[5] *See* Kate Knibbs, *Big Data and the Death of Internet Privacy*, http://www.mobiledia.com/news/154591.html (last visited Apr. 21, 2014).

[6] *See* Herb Weisman, *When Big Data Becomes Big Brother,* http://www.thefiscaltimes.com/Articles/2014/04/09/When-Big-Data-Becomes-Big-Brother (last visited Apr. 21, 2014).

recognition? Distributed computing? A specific technology that can be used to draw insights from data? Or just generally the impact mass information collection and storage is having on society?"[7] My answer is "Yes, to all of the above." But not pre-crime, I hope. For our purposes here, I'll rely on one common definition. Big data is data that has three characteristics:

- First, the data has large **volume** – there is a lot of it. And I mean a LOT. As of 2012, when Facebook had 15% fewer users than today, its largest Hadoop cluster had over 100 petabytes of data.[8] That's equivalent to 200,000 years of digital music, which I think is about six Pink Floyd songs.
- Second, the data has significant **variety**. That is, the data may be structured (like an Excel table with column headers), semi-structured (like an Excel table without column headers), or unstructured (like a folder full of electronic documents and images).
- Third, the data has high **velocity**. Meaning that data is being produced and analyzed at a rapid rate, often in real time.

Experts also use "big data" to refer to the set of software tools and techniques for collecting, storing, and analyzing data with these three characteristics.

Big data isn't completely new. Large companies like Wal-mart have been collecting and analyzing terabytes of consumer data since the early 1990s.[9] However, today the tools for big data analysis are cost-effective even for small companies. Additionally, there is a lot more data, of many different kinds, being produced today.[10] Back then, Wal-mart was collecting data on what consumers purchased from its stores and batch processing it each night. Today, a big data

---

[7] Michael Sherman, *The Death of 'Big Data'*, http://www.texasenterprise.utexas.edu/2014/02/27/innovation/death-big-data (last visited Apr. 21, 2014).

[8] Andrew Ryan, *Under the Hood: Hadoop Distributed Filesystem reliability with Namenode and Avatarnode*, https://www.facebook.com/notes/facebook-engineering/under-the-hood-hadoop-distributed-filesystem-reliability-with-namenode-and-avata/10150888759153920 (last visited Apr. 21, 2014); Ben Foster, *How Many Users on Facebook,* http://www.benphoster.com/facebook-user-growth-chart-2004-2010/ (last visited Apr. 21, 2014).

[9] *See* Thomas Wailgum, *45 Years of Wal-Mart History: A Technology Time Line*, http://www.cio.com/article/147005/45_Years_of_Wal_Mart_History_A_Technology_Time_Line (last visited Apr. 21, 2014).

[10] According to Intel, in 2015 the world will produce 8.1 zettabytes of data – 1,500 times more than all the data produced from the beginning of time until 2003. *See* Intel, Big Data 101 Video, *available at* http://www.intel.com/content/www/us/en/big-data/big-data-analytics-turning-big-data-into-intelligence-cmpg.html.

project might combine such purchase data with shipping, inventory, and even traffic data to achieve just-in-time delivery of consumer products.

Benefits of Big Data. This change in tools and data sources has great potential to make our lives better. As Professor Sinan Aral of New York University has explained, "Revolutions in science have often been preceded by revolutions in measurement."[11] The promise is that big data techniques will extract from data knowledge that will help us better understand the world, similar to how the microscope's magnification of tiny things led to the germ theory of disease. Today we already benefit from Amazon, e-Bay, Netflix, and many other online merchants' use of big data to generate customized user recommendations. Big data is used today to aggregate millions of GPS signals to predict commute times, to identify potential causes of disease, and to detect and prevent credit card fraud. Kaiser Permanente used big data analysis to discover an increased chance of heart attack or cardiac death among users of Vioxx as compared to users of a competing medication.[12] Scientists are using massive data sets and powerful analytic tools to make progress on many of the most difficult problems in the health sciences and hard sciences. And many new uses are emerging, particularly because consumers are no longer simply data points to be researched. Today's consumers are themselves producers and consumers of big data, whether posting billions of cat photos on Facebook, using Bing's flight price predictors to make travel plans, or joining the self-quantification movement by wearing a FitBit Flex and using a Withings [Wye-theengs] bathroom scale. As more of our everyday existence becomes measurable and recordable, the more potential there is for big data to provide helpful insights.

---

[11] *Data, Data Everywhere,* THE ECONOMIST (Feb. 25, 2014) *available at* http://www.economist.com/node/15557443.

[12] *See* Rachael King, *Data Helps Drive Lower Mortality Rate at Kaiser*, http://blogs.wsj.com/cio/2013/12/05/data-helps-drive-lower-mortality-rate-at-kaiser/ (last visited Apr. 21, 2014).

Technical Challenges. Some advocates of big data go even further, asserting that big data will change how we approach science. Specifically, some say that with big enough data sets, we can simply looks for correlations between variables or sets of variables without needing a theory for why the two variables might be related or how the causation might work.[13] If this were true, it would make many intractable problems seem more approachable.

These claims are challenged by some data scientists, however. David Spiegelhalter, Winton Professor of the Public Understanding of Risk at Cambridge University, is one of the skeptics. He has said, "There are a lot of small data problems that occur in big data… They don't disappear because you've got lots of the stuff. They get worse."[14] In particular, there are two technical concerns regarding big data that I'd like to discuss today. First, there are so-called "signal problems," where the data set, huge as it may be, is not representative of the real world. Kate Crawford describes the City of Boston's StreetBump mobile app as an example of this kind of problem.[15] The StreetBump app monitors GPS and accelerometer data on users' phones to passively detect potholes and report them to the city. However, the data is noticeably tilted toward finding potholes in areas where a higher percentage of the driving population owns a smartphone. Thus, because the underlying data did not accurately reflect the real world, neither did the result of the analysis.

Second, big data is susceptible to the "multiple comparisons problem." Big data tools are particularly good at discovering correlations in complex data sets. However, as a recent New

---

[13] *See generally* PATRICK TUCKER, THE NAKED FUTURE: WHAT HAPPENS IN A WORLD THAT ANTICIPATES YOUR EVERY MOVE? (Mar. 2014).

[14] Tim Harford, *Big data: are we making a big mistake?,* http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz2xGZdY1so (last visited Apr. 21, 2014).

[15] Kate Crawford, *The Hidden Biases in Big Data*, http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/ (last visited Apr. 21, 2014).

York Times op-ed pointed out, big data can't tell us which correlations are important and which are spurious.[16] If a scientist examines a single data set for 100 different correlations, probability says he will find five patterns that appear statistically significant but which are a result of random chance. This problem actually gets worse in larger data sets because there are more possible correlations to test. This may be an even more significant problem when the investigator is simply exploring a big data set without a particular question in mind. In such cases, it is easy to find "statistically significant" correlations that are actually the result of pure chance.

Both of these problems are reminders that data, even big data, isn't knowledge or wisdom. It can be misleading. Even worse, data-driven decisions can *seem* right while being wrong. Political polling expert Nate Silver notes that "[o]ne of the pervasive risks that we face in the information age … is that even if the amount of knowledge in the world is increasing, the gap between *what we know* and what *we think* we know may be widening."[17]

The good news is that these problems are not new. Statisticians have been developing methods to deal with bias and sample errors for as long as there have been statisticians. Data scientists will need to update those long-standing techniques to correct these problems in the context of big data. These problems do not negate the significant potential benefits of big data techniques. But they do mean that big data analysis is not an all-powerful technique. It is a tool that has certain limitations, and like all tools, it can be used or misused. Both big data boosters and big data skeptics should pay attention to these limitations. By pulling some of the hype out of the debate, we can better ensure an appropriate and proportional response.

---

[16] Gary Marcus and Ernest Davis, *Eight (No, Nine!) Problems with Big Data,* THE NEW YORK TIMES (Apr. 6, 2014) *available at* http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html.

[17] NATE SILVER, THE SIGNAL AND THE NOISE: WHY SO MANY PREDICTIONS FAIL – BUT SOME DON'T (2012).

III. Privacy and Other Concerns

Like many new technologies, big data raises concerns about how current laws will protect consumers. Of course, many types of big data research have nothing to do with individuals and do not raise these concerns. However, some consumer and privacy advocates are concerned that consumers will suffer harm from other uses of big data. Such advocates are particularly uncomfortable about the implications of large, persistent data sets containing information on individual customers. These concerns generally fall into three categories. First, some of the concerns about big data apply to data more generally, and the FTC has been actively addressing these issues for years. Second, big data does raise some genuinely new challenges, particularly about how we can adapt the FIPPs framework to work with big data.[18] These issues need further research and careful consideration by stakeholders. Third, there are concerns over fairness and discrimination in big data. While these aren't really privacy issues as such, they are important and worth studying carefully. I would like to explore these three points more fully.

Many concerns over big data are not unique to big data. Many big data concerns are also concerns for traditional "small data" and are already familiar to the FTC. For example, without adequate security safeguards, any data, big or small, can fall into the wrong hands. Recent reports of data breaches at retailers and other businesses obviously raise serious concerns. Yet there are real market and reputational incentives for companies to get data security right in the big data context. Furthermore, the FTC has for years been actively enforcing basic data security requirements to address consumer harm and has brought more than fifty data security cases. Recently, in *FTC v. Wyndham*, a federal district court confirmed that the FTC has

[18] *See* DEPT. OF COMMERCE INTERNET POLICY TASK FORCE, COMMERCIAL DATA PRIVACY AND INNOVATION IN THE INTERNET ECONOMY: A DYNAMIC POLICY FRAMEWORK, at 11 (2010) (describing the 1973 origin of the Fair Information Privacy Practices framework at the Department of Health, Education, and Welfare), *available at* http://www.ntia.doc.gov/files/ntia/publications/iptf_privacy_greenpaper_12162010.pdf.

authority to protect consumers from unfair data security practices by bringing such cases.[19] And while some recent "big data" breaches are very large in scale, this is not a new development. For example, in 2009 the FTC investigated a data breach at Heartland Payment Systems, where hackers stole more than 130 million credit card numbers.[20] The FTC's data security enforcement framework is not perfect; I would like to develop more concrete guidance to industry, for example. But I haven't seen anything that suggests that big data technology raises fundamentally new data security issues.

Similarly, some groups also argue that certain types of particularly sensitive data, such as data about children, health, or finances, deserve heightened protection when stored in big data sets. Of course, the FTC already recognizes the need to more thoroughly protect such types of data, whether the data is in big data or small data environments.

Tension with Certain Fair Information Practice Principles. Other concerns about big data do appear to raise new issues, however. In particular, maximizing the benefits of big data may create tension with the notice and the purpose limitation principles in the FIPPs. These two related principles say that an information collector should inform consumers about the collection and its purpose and get the consumer's consent for the collection for that purpose. Yet much of the promise of big data is that it can find something new and useful in the data that could not have been anticipated at the time of collection. But companies cannot give notice at the time of collection for unanticipated uses. Furthermore, in many cases, data scientists create one big data set from many other smaller collections that initially served different purposes and may have

[19] *FTC v. Wyndham Worldwide Corp.*, No. 13-1887, 2014 U.S. Dist. LEXIS 47622 (D.N.J. Apr. 7, 2014).

[20] Robert McMillan, "SEC, FTC investigating Heartland after data theft," Feb. 25, 2009, COMPUTERWORLD, *available at* http://www.computerworld.com/s/article/9128658/SEC_FTC_investigating_Heartland_after_data_theft.

been collected at different times from a wide range of sources. As such, it is difficult or impossible to notify individuals of the new purpose for which the data is being used.

The FIPPs principle of data minimization is also in tension with the incentives of big data. Part of the promise of big data is to pull knowledge from data points whose value was previously unknown. Thus, retention of as much data as possible for lengthy amounts of time is a common practice. Strictly limiting the collection of data to the particular task currently at hand and disposing of it afterwards would handicap the data scientist's ability to find new information to address future tasks. Certain de-identification techniques such as anonymization, although not perfect, can help mitigate some of the risks of comprehensive data retention while permitting innovative big data analysis to proceed.

I believe FIPPs remains a solid framework and is flexible enough to accommodate a robust big data industry, but we have some work to do to resolve these tensions. I welcome your ideas on how we can do this.

Other, Non-Privacy Concerns. Finally, some advocates worry that companies will use big data techniques to prejudge or discriminate against individuals unfairly or erroneously without recourse. The concern is that a researcher could collect non-sensitive information about a consumer and then use big data analysis to infer certain sensitive characteristics about that consumer.

This is a complicated issue that we need to know more about. First, companies have long engaged in this type of consumer targeting with more traditional tools. It is not clear how much additional value big data analysis will bring, because, as noted earlier, big data analysis is not a foolproof tool for all questions. Second, it is not yet clear how likely companies are to use such

an approach. Third, if companies do engage in this sort of analysis, we need to determine how they might use such information.

This third point, the type of use, matters, as our legal framework restricts certain uses of data regardless of how it was collected. Specifically, the Fair Credit Reporting Act establishes constraints for companies that make certain uses of data: creditworthiness, insurance eligibility, evaluation for employment, and renter background checks. Passed in 1970 in response to the creation of credit reporting bureaus, the FCRA could be considered the first "big data" bill. In fact, the FTC has applied the FCRA in a "big data" context. In 2012, the FTC entered into an $800,000 settlement with Spokeo, a company that assembles personal profiles of individuals from information in public records, white pages, and social networking sites.[21] Spokeo was allegedly marketing personal information to potential employers in violation of the FCRA. I believe the FCRA may provide a useful model for the types of big data uses that raise significant consumer concern. Any new exploration of FCRA-like use restrictions, however, should not undermine the continued application of many of the FIPPs principles, which have worked well for decades. But I hope we can explore whether specifically prohibiting certain clearly impermissible uses could help protect consumers while enabling continued innovation in big data. Any exploration of this FCRA-like approach should involve a detailed cost-benefit analysis, of course.

None of this is to denigrate the establishment of principles to guide the collection of data. Such principles can and do serve as important best practices or industry standards. That is why I have repeatedly supported as best practices many (although not all) of the recommendations of

[21] Press Release, Fed. Trade Comm'n, Spokeo to Pay $800,000 to Settle FTC Charges Company Allegedly Marketed Information to Employers and Recruiters in Violation of FCRA (June 12, 2012) *available at* http://www.ftc.gov/news-events/press-releases/2012/06/spokeo-pay-800000-settle-ftc-charges-company-allegedly-marketed.

the FTC's 2012 report on "Protecting Consumer Privacy in an Era of Rapid Change."[22] Some of the most relevant recommendations of that report for big data include:

- **Privacy by Design** – Companies should build in consumer privacy protections at every stage in developing their products. These protections include reasonable security for consumer data and reasonable procedures to promote data accuracy. In the big data context, built-in de-identification measures could play an important role in protecting consumer privacy.
- **Simplified Choice for Businesses and Consumers** – Recognizing that there is no single best way to offer notice and choice in all circumstances, companies should adopt notice and choice options that appropriately reflect the context of the transaction or the relationship the company has with the consumer. In the big data context, this may be challenging, but I believe it is a principle worth continuing to pursue.
- **Greater Transparency** – Companies should disclose details about their collection and use of consumers' information and provide consumers access to the data collected about them.

I believe these best practices are flexible enough to remain useful in many, if not all, situations. Companies that embrace these principles would benefit their customers. Of course, best practices necessarily change with the environment. We must work to determine what changes may be necessary to protect and advance consumer welfare.

IV. FTC's Role in Big Data

The FTC can help ensure that the promise of big data is realized by using our unique set of enforcement and policy tools. First, the FTC is an enforcement agency and it can and should use its traditional deception and unfairness authority to stop consumer harms that may arise from the misuse of big data. Strong enforcement will help not only consumers but also other companies using big data analysis by policing actors that may tarnish the technology itself. Second, we can use our convening power and our policy and R&D functions to better understand big data technology; the new business models it may enable; the applicability of existing

[22] FED. TRADE COMM'N, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE: RECOMMENDATIONS FOR BUSINESSES AND POLICYMAKERS (Mar. 2012) *available at* http://www.ftc.gov/reports/protecting-consumer-privacy-era-rapid-change-recommendations-businesses-policymakers.

regulatory structures, including self-regulation; market dynamics; and the nature and extent of likely consumer and competitive benefits and risks.

I am happy to say that the FTC is working hard to understand the promise and risks of big data and related technologies. Last year the FTC held a workshop on the Internet of Things in which we explored both the potential benefits and risks to consumers of this new environment of constant data flow.[23] More recently, the FTC hosted a workshop on alternative scoring mechanisms to evaluate the potential implications of new types of scoring that rely on big data predictive analytics to provide identity verification, fraud prevention, and marketing and other services.[24] In May, the Commission will host an event on consumer generated and controlled health data – one of the newest and most interesting sources of big data.[25] The FTC also announced a September big data workshop that will explore some of the benefits and potential risks of various big data techniques.[26] I hope that many of you will attend and contribute to these upcoming events. Your insights will help educate the FTC on big data technology, the key issues it raises, and the FTC's proper course of action in this area.

As the FTC uses these various institutional tools to engage with big data issues, two principles should guide our work. First, as with all dynamic markets, we must approach big data technologies with what I call regulatory humility. Our most successful technological advances,

---

[23] Press Release, Fed. Trade Comm'n, FTC Announces Agenda, Panelists for Upcoming Internet of Things Workshop (Nov. 8, 2013) *available at* http://www.ftc.gov/news-events/press-releases/2013/11/ftc-announces-agenda-panelists-upcoming-internet-things-workshop.

[24] Press Release, Fed. Trade Comm'n, FTC Announces Agenda, Panelists for Alternative Scoring Seminar (Mar. 14, 2014) *available at* http://www.ftc.gov/news-events/press-releases/2014/03/ftc-announces-agenda-panelists-alternative-scoring-seminar.

[25] Press Release, Fed. Trade Comm'n, FTC to Host Spring Seminars on Emerging Consumer Privacy Issues (Dec. 2, 2013) *available at* http://www.ftc.gov/news-events/press-releases/2013/12/ftc-host-spring-seminars-emerging-consumer-privacy-issues.

[26] Press Release, Fed. Trade Comm'n, FTC to Examine Effects of Big Data on Low Income and Underserved Consumers at September Workshop (Apr. 11, 2014) *available at* http://www.ftc.gov/news-events/press-releases/2014/04/ftc-examine-effects-big-data-low-income-underserved-consumers.

such as the Internet itself, have generated massive amounts of consumer welfare and have thrived largely because market participants have enjoyed wide latitude to experiment with new technology-driven business models, allowing the market to determine which of those models succeeds or fails. This is the right approach. Second, we must identify substantial consumer harm before taking action. Thus, the FTC should remain vigilant for deceptive and unfair uses of big data, but should avoid preemptive action that could preclude entire future industries. Ultimately, our work as an agency should help strengthen competition and the market to better provide beneficial outcomes in response to consumer demand, rather than to try to dictate desired outcomes to the market.

V. Conclusion

To conclude, big data is not just coming – it is here. It's neither a miracle cure nor a plague. It is a powerful tool with great promise and some risks. Many of the concerns raised by big data are suitably handled by current law and policy. But where there are new issues, regulators need to work with people like you to understand the issues deeply and focus our enforcement actions on situations where improper use of consumer information causes substantial harm. This approach will free entrepreneurs to innovate with big data tools while simultaneously helping to ensure that consumer privacy remains protected.

Thank you for having me today, and I look forward to your questions.