Big Data for All: Privacy and User Control in the Age of Analytics

***Omer Tene*[1] *and Jules Polonetsky*[2]**

[1] Associate Professor, College of Management Haim Striks School of Law, Israel; Senior Fellow, Future of Privacy Forum; Visiting Researcher, Berkeley Center for Law and Technology; Affiliate Scholar, Stanford Center for Internet and Society. I would like to thank the College of Management Haim Striks School of Law research fund and the College of Management Academic Studies research grant for supporting research for this article.

[2] Co-chair and Director, Future of Privacy Forum.

Abstract

We live in an age of "big data". Data have become the raw material of production, a new source for immense economic and social value. Advances in data mining and analytics and the massive increase in computing power and data storage capacity have expanded by orders of magnitude the scope of information available for businesses and government. Data are now available for analysis in raw form, escaping the confines of structured databases and enhancing researchers' abilities to identify correlations and conceive of new, unanticipated uses for existing information. In addition, the increasing number of people, devices, and sensors that are now connected by digital networks has revolutionized the ability to generate, communicate, share, and access data. Data creates enormous value for the world economy, driving innovation, productivity, efficiency and growth. At the same time, the "data deluge" presents privacy concerns which could stir a regulatory backlash dampening the data economy and stifling innovation. In order to craft a balance between beneficial uses of data and in individual privacy, policymakers must address some of the most fundamental concepts of privacy law, including the definition of "personally identifiable information", the role of individual control, and the principles of data minimization and purpose limitation. This article emphasizes the importance of providing individuals with access to their data in usable format. This will let individuals share the wealth created by their information and incentivize developers to offer user-side features and applications harnessing the value of big data. Where individual access to data is impracticable, data are likely to be de-identified to an extent sufficient to diminish privacy concerns. In addition, organizations should be required to disclose their decisional criteria, since in a big data world it is often not the data but rather the inferences drawn from them that give cause for concern.

Introduction

Big data is upon us.[3] Over the past few years, the volume of data collected and stored by business and government organizations has exploded.[4] The trend is driven by reduced costs of storing information and moving it around in conjunction with increased capacity to instantly analyze heaps of unstructured data using modern experimental methods, observational and longitudinal studies and large scale simulations.[5] Data are generated from online transactions, email, video, images, clickstream, logs, search queries, health records and social networking interactions; gleaned from increasingly pervasive sensors deployed in infrastructure such as communications networks, electric grids, global positioning satellites, roads and bridges,[6] as well as in homes, clothing, and mobile phones.[7]

The Obama Administration has recently announced a new, multi-agency big data research and development initiative aimed at advancing the core scientific and technological means of managing, analyzing, visualizing and extracting information from large, diverse, distributed, and heterogeneous data sets.[8] This is based on recognition of the immense social and economic value captured in information and the intention to unleash it in order to progress from data to knowledge to action.[9] Big data boosts the economy, transforming traditional business models and creating new opportunities through the use of business intelligence, sentiment analysis and analytics; advances scientific research, transforming scientific methods from hypothesis-driven

---

[3] Steve Lohr, The Age of Big Data, NY TIMES, February 11, 2012, http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all; Steve Lohr, How Big Data Became So Big, NY TIMES, August 11, 2012, http://www.nytimes.com/2012/08/12/business/how-big-data-became-so-big-unboxed.html; Janna Anderson & Lee Rainie, The Future of Big Data, PEW INTERNET & AMERICAN LIFE PROJECT, July 20, 2012, http://pewinternet.org/~/media//Files/Reports/2012/PIP_Future_of_Internet_2012_Big_Data.pdf.

[4] Kenneth Cukier, Data, data everywhere, THE ECONOMIST, February 25, 2010, http://www.economist.com/node/15557443; for some examples *see* World Economic Forum, Big Data, Big Impact: New Possibilities for International Development (2012), http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf.

[5] *See, e.g.*, Trevor Hastie, Robert Tibshirani & Jerome Friedman, THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE, AND PREDICTION (Springer Verlag, New York, 2009).

[6] For the erosion of privacy in the public sphere *see* UNITED STATES V. JONES, 132 S.Ct. 945 (2012).

[7] Omer Tene, Privacy: The New Generations, 1 INT'L DATA PRIVACY LAW 15 (2011), http://idpl.oxfordjournals.org/content/1/1/15.full.

[8] Office of Science and Technology Policy, Executive Office of the President, News Release, Obama Administration Unveils "Big Data" Initiative: Announces $200 Million in New R&D Investments, March 29, 2012, http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf.

[9] World Economic Forum, Personal Data: The Emergence of a New Asset Class, 2011, http://www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf; Steve Lohr, New U.S. Research Will Aim at Flood of Digital Data, NY TIMES, March 29, 2012, http://www.nytimes.com/2012/03/29/technology/new-us-research-will-aim-at-flood-of-digital-data.html?_r=2.

to data-driven discovery;[10] and furthers national goals such as optimization of natural resources, response to national disasters and enhancement of critical information infrastructure.[11]

Nevertheless, the extraordinary societal benefits of big data, including breakthroughs in medicine, data security, and energy use, must be reconciled with increased risks to individuals' privacy.[12] As is often the case, the technological and business developments on the ground have far outpaced the existing legal frameworks, which date back from an era of mainframe computers, predating the Internet, mobile and cloud computing.[13] For the past four decades, the tension between data innovation and informational privacy has been moderated by a set of principles broadly referred to as the "FIPPs", based on a framework set in the 1980 OECD Guidelines.[14] In their latest version presented by the White House this year, the FIPPs include the principles of individual control; transparency; respect for context; security; access and accuracy; focused collection; and accountability.[15] The big data paradigm challenges some of these fundamental principles, including the scope of the framework (often addressed by framing the term "personally identifiable information" (PII)); the concepts of data minimization

[10] *See* Chris Anderson, The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, WIRED, June 23, 2008, http://www.wired.com/science/discoveries/magazine/16-07/pb_theory; *also see* presentation of Peter Norvig, The Unreasonable Effectiveness of Data, September 23, 2011, http://www.youtube.com/watch?v=yvDCzhbjYWs.

[11] Farnam Jahanian, Assistant Director, National Science Foundation, NSF Keynote at TechAmerica's Big Data Congressional Briefing, May 2, 2012, http://www.youtube.com/watch?v=Do_IPa6-E9M.

[12] Omer Tene & Jules Polonetsky, Privacy in the Age of Big Data: A Time for Big Decisions, 64 STAN. L. REV. ONLINE 63 (2012).

[13] *See* OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, ORG. FOR ECON. CO-OPERATION & DEV. (September 23, 1980) [hereinafter: OECD Guidelines], http://www.oecd.org/document/18/0,3343,en_2649_34255_1815186_1_1_1_1,00.html; Council of Europe Convention 108 for the Protection of Individuals with regard to Automatic Processing of Personal Data, Strasbourg, 28 January, 1981, http://conventions.coe.int/treaty/en/treaties/html/108.htm; Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, 1995 O.J. (L 281) 31 (Nov. 23, 1995) [hereinafter: European Data Protection Directive], http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:1995:281:0031:0050:EN:PDF; and in the United States: The Privacy Act of 1974, Pub. L. No. 93-579, 88 Stat. 1897 (December 31, 1974). All of the major frameworks are being reviewed this year. *See* 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 [hereinafter: White House Blueprint] http://www.whitehouse.gov/sites/default/files/privacy-final.pdf; Federal Trade Commission Report, Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers, March 2012 [hereinafter: FTC Final Report], http://ftc.gov/os/2012/03/120326privacyreport.pdf; Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), Brussels, January 25, 2012, COM(2012) 11 final [hereinafter: European Data Protection Regulation], http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf.

[14] The OECD Guidelines include the principles of collection limitation; data quality; purpose specification; use limitation; security safeguards; openness; individual participation; and accountability. OECD Guidelines, *ibid.*

[15] White House Blueprint, *ibid*.

("focused collection") and consent ("individual control" and "respect for context"); and the right of individual access ("access and accuracy").[16]

This article addresses the legal issues arising from the big data debate. It suggests that the FIPPs should be viewed as a set of levers that must be adjusted to adapt to varying business and technological conditions. Indeed, the ingenuity of the FIPPs, which has made them resilient to momentous change, is in their flexibility: some principles retract while others expand depending on the circumstances. In the context of big data, this means relaxing data minimization and consent requirements while emphasizing transparency, access and accuracy. The shift is from empowering individuals at the point of information collection, which traditionally revolved around opting into or out of seldom read, much less understood corporate privacy policies; to allowing them to engage with and benefit from information already collected, thereby harnessing big data for their own personal usage. Further, such exposure will prevent the existence of "secret" databases and leverage societal pressure to constrain any unacceptable uses.

The article assesses the definition of PII in a world where de-identification is often reversible and sometimes detrimental to the integrity of the very data it aims to protect. It seeks to reconcile the current technological and business realities with the data minimization and purpose limitation principles. These principles are antithetical to big data, which is premised on data maximization – the more data processed the finer the conclusions – and seeks to uncover surprising, unanticipated correlations.

The article suggests that to solve the big data privacy quandary, individuals must be offered meaningful rights to access their data in usable, machine-readable format. This, in turn, will unleash a wave of innovation for user-side applications and services based on access to personal information, a process we refer to as the "featurization" of big data. Featurization will allow individuals to declare their own policies, preferences and terms of engagement, and do it in ways that can be automated both for them and for the companies they engage.[17] Where individual access to data is impracticable, data are likely to be de-identified to an extent sufficient to diminish privacy concerns. Where access is possible, organizations must provide it with robust mechanisms for user authentication and through secure channels to prevent leakage. This implies the development of user-centric or federated identity management schemes, which include single sign-on capability and at the same time do not become vehicles for universal surveillance.[18]

---

[16] Federal Trade Commission Commissioner Julie Brill recently said: "Big Data's impact on privacy is requiring some new and hard thinking by all of us." Comments in Big Data, Big Issues Conference, Fordham University School of Law, March 2, 2012, http://ftc.gov/speeches/brill/120228fordhamlawschool.pdf.

[17] *See* Doc Searls, The Customer as a God, WALL STREET JOURNAL, July 20, 2012, http://online.wsj.com/article/SB10000872396390444873204577535352521092154.html.

[18] *See, e.g.*, Ann Cavoukian, 7 Laws Of Identity: The Case for Privacy-Embedded Laws Of Identity in the Digital Age, 2006, http://www.identityblog.com/wp-content/resources/7_laws_whitepaper.pdf.

To minimize concerns of untoward data usage, organizations should disclose the logic underlying their decision-making processes to the extent possible without compromising their trade secrets or intellectual property rights. As danah boyd and Kate Crawford recently noted: "In reality, working with Big Data is still subjective, and what it quantifies does not necessarily have a closer claim on objective truth…".[19] It is imperative that individuals have insight into the decisional criteria of organizations lest they face a Kafkaesque machinery that manipulates lives based on opaque justifications. While we recognize the practical difficulties of mandating disclosure without compromising organizations' "secret sauce", we trust that a distinction can be drawn between proprietary algorithms, which would remain secret, and decisional criteria, which would be disclosed.

In Part One, we describe some of the benefits of big data to individuals and society at large, including medical research, smart grid and traffic management. Some of the use cases are so compelling that few would argue they should be forgone in light of the incremental risk to individuals' privacy. In Part Two, we lay out some of the risks of big data, including the unidirectional, incremental chipping away at informational privacy; the social stratification exacerbated by predictive analysis; and the exclusion of individuals from the value generated by their own information. In Part Three, we address big data's challenges to existing privacy rules, including the definition of PII; the principle of data minimization; and the concept of meaningful, informed consent. Part Four makes the case for providing individuals with useful access to their data, allowing them to share the gains generated by the combination of their information with resources invested by businesses and government; and for requiring organizations to be transparent with respect to the decisional criteria underlying their big data choices.

Big Data: Big Benefits

Big data is a big industry. Research conducted at the Massachusetts Institute of Technology shows that companies that use "data-directed decision-making" enjoy a 5-6% increase in productivity.[20] There is a strong link between effective data management strategy and financial performance. Companies that use data most effectively stand out from the rest. A report by the McKinsey Global Institute (MGI) demonstrates the transformative effect that big data has had on entire sectors ranging from health care to retail to manufacturing to political campaigns.[21]

---

[19] danah boyd & Kate Crawford, Critical Questions for Big Data, INFO. COMM. & SOC'Y (MAY 2012), at p. 6.

[20] Erik Brynjolfsson, Lorin Hitt and Heekyung Kim, Strength in Numbers: How Does Data-Driven Decision-Making Affect Firm Performance?, April 2011, http://www.a51.nl/storage/pdf/SSRN_id1819486.pdf; also *see* recent WEF report referring to personal data as "the new oil", a new asset class emerging as the most valuable resource of the 21st century, *supra* note 9.

[21] McKinsey Global Institute, Big data: The next frontier for innovation, competition, and productivity, June 2011 ("MGI Report"), http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_fron

Just as it helps businesses increase productivity, big data allows governments to improve public sector administration and assists global organizations in analyzing information to devise strategic planning. Demand for big data is accelerating. MGI projected that the United States already needs 140,000 to 190,000 more workers with "deep analytical" expertise and 1.5 million more data-literate managers.[22]

In this chapter, we present some anecdotal examples of the benefits that big data has brought. When considering the risks that big data poses to individuals' privacy, policymakers should be minded of its sizable benefits. Privacy impact assessments (PIA), systematic processes undertaken by government and business organizations to evaluate the potential risks to privacy of products, projects or schemes, often fail to bring these benefits into account.[23] Concluding that a project raises privacy risks is not sufficient to discredit it. Privacy risks must be weighed against non-privacy rewards. And while numerous mechanisms exist to assess privacy risks,[24] we still lack a formula to work out the balance.[25]

At the same time, under existing market conditions, the benefits of big data do not always (some say, ever) accrue to the individuals whose personal data are collected and harvested.[26] This creates a twofold problem: on the one hand, individuals should not be required to volunteer their information with little benefit beyond feeding voracious corporate appetites; on the other hand, self interest should not frustrate societal values and benefits such as law enforcement, public health, or economic progress. If individuals could reap some of the gains of big data, they would be incentivized to actively participate in the data economy, aligning their own self-interest with broader societal goals.

---

tier_for_innovation; Thomas Edsall, Let the Nanotargeting Begin, NY TIMES CAMPAIGNSTOPS BLOG, April 15, 2012, http://campaignstops.blogs.nytimes.com/2012/04/15/let-the-nanotargeting-begin.

[22] Ben Rooney, Big Data's Big Problem: Little Talent, WALL STREET JOURNAL TECH EUROPE, April 26, 2012, http://blogs.wsj.com/tech-europe/2012/04/26/big-datas-big-problem-little-talent/?mod=google_news_blog.

[23] E-Government Act of 2002, sec. 208, 44 U.S.C. § 101, Pub. L. No. 107–347; European Data Protection Regulation, art. 33-34.

[24] *See, e.g.*, Roger Clarke, An Evaluation of Privacy Impact Assessment Guidance Documents, 1 INT'L DATA PRIVACY LAW 111 (2011); United States Securities and Exchange Commission, PRIVACY IMPACT ASSESSMENT (PIA) GUIDE, January 2007, http://www.sec.gov/about/privacy/piaguide.pdf; United States Department of Homeland Security, PRIVACY IMPACT ASSESSMENTS, THE PRIVACY OFFICE OFFICIAL GUIDANCE, June 2010, http://www.dhs.gov/xlibrary/assets/privacy/privacy_pia_guidance_june2010.pdf; United States Department of Justice, Office of Privacy and Civil Liberties, PRIVACY IMPACT ASSESSMENTS, OFFICIAL GUIDANCE, August 2010, http://www.justice.gov/opcl/pia_manual.pdf.

[25] For example, if analysis of de-identified online search engine logs enabled identification of a life threatening epidemic in *x*% of cases thus saving *y* lives; should such analysis be permitted assuming a z% chance of re-identification of a certain subset of search engine users? This is a meta-privacy question, which must be answered by policymakers implementing more than just a PIA; the PIA only solves one side of the equation.

[26] *See, e.g.*, Natasha Singer, Consumer Data, but Not for Consumers, NY TIMES, July 21, 2012, http://www.nytimes.com/2012/07/22/business/acxiom-consumer-data-often-unavailable-to-consumers.html.

Healthcare

Dr. Russ Altman, a professor of medicine and bioengineering at Stanford, and his colleagues made a groundbreaking discovery last year. They found that when taken together, Paxil, the blockbuster antidepressant prescribed to millions of Americans, and Pravachol, a highly popular cholesterol-reducing drug, have a dreadful side effect. They increase patients' blood glucose to diabetic levels. Each drug taken alone does not have the diabetic side effects; hence the Food and Drug Administration (FDA) approved the drugs for use. Surely the FDA cannot test each and every drug for every conceivable interaction.

Altman and his team made their discovery by pursuing statistical analysis and data mining techniques to identify patterns in large datasets. They analyzed information in the Adverse Event Reporting System (AERS), a database maintained by the FDA to collect adverse drug event reports from clinicians, patients, and drug companies for more than 30 years.[27] Using the AERS, they created a "symptomatic footprint" for diabetes-inducing drugs (*i.e.*, the side effects a patient might report if she had undiagnosed diabetes), and then searched for that footprint in interactions between pairs of drugs not known to induce such effects when taken alone. Four pairs of drugs were found to leave the footprint, the most commonly prescribed ones being Paxil and Pravachol.

Next, the scientists approached Microsoft Research to examine de-identified Bing search engine logs, querying whether a higher proportion of users who searched for *both* "Paxil" and "Pravachol" also typed in words related to the "symptomatic footprint" (such as "headache" or "fatigue") than those who searched for just Paxil or Pravachol *separately*. Sure enough, their research hypothesis found support in that big data set too. Users who searched Bing for the name of both drugs together were much likelier to search for diabetes-related side effects than users who searched for only one of the drugs.[28]

By implementing a novel signal detection algorithm that identifies statistically significant correlations, the researchers were thus able to parse out latent adverse effect signals from spontaneous reporting systems.[29] Nearly 15 million Americans use either Paxil or Pravachol; an estimated one million used to take both. For these users, the work of Altman and his colleagues was potentially life-saving.[30]

---

[27] Adverse Event Reporting System (AERS), http://www.fda.gov/drugs/guidancecomplianceregulatoryinformation/surveillance/adversedrugeffects/default.htm.

[28] Watch Altman describe the research process, including the search engine logs analysis, in the Stanford Law Review Online 2012 Symposium, The Privacy Paradox: Privacy and Its Conflicting Values, February 3, 2012, http://www.stanfordlawreview.org/symposia (from minute 32 of the video).

[29] *Also see* David Reshef, Yakir Reshef *et al*, Detecting Novel Associations in Large Data Sets, 334 SCIENCE 1518 (December 2011).

[30] Nicholas Tatonetti, Guy Haskin Fernald & Russ Altman, A Novel Signal Detection Algorithm for Identifying Hidden Drug-Drug Interactions in Adverse Event Reports, J. AM. MED. INFORM. ASSOC. (2011).

The findings of Altman and his team are not the sole major breakthrough based on big data analysis in healthcare. The discovery of Vioxx's adverse drug effects, which led to its withdrawal from the market, was made possible by analysis of clinical and cost data collected by Kaiser Permanente, the California-based managed-care consortium.[31] Had Kaiser Permanente not aggregated clinical and cost data, researchers might not have been able to attribute 27,000 cardiac arrest deaths occurring between 1999 and 2003 to use of the drug. Similarly, researchers in South Africa discovered a positive relationship between therapeutic vitamin B use and delay of progression to AIDS and death in HIV-positive patients.[32] This was a critical finding at a time and in a region where therapies for people living with HIV are well beyond the financial means of most patients. The researchers noted that "[n]onlinear statistical analysis … can help elucidate clinically-relevant relationships within a large patient population such as observational databases."[33] Another oft-cited example is Google Flu Trends, which predicts and locates outbreaks of the flu making use of information – aggregate search queries – not originally collected with this innovative application in mind. Of course, early detection of disease activity, when followed by rapid response, can reduce the impact of both seasonal and pandemic influenza.[34]

A further example is the National Retail Data Monitor (NRDM), which keeps tabs on sales of over-the-counter healthcare items from 21,000 outlets across the United States. By analyzing what remedies people are buying, health officials can anticipate short-term trends in illness transmission. Data from the NRDM show that sales of over-the-counter products like cough medicine and electrolytes spike before visits to the emergency room do, and that the lead-time can be significant – two and a half weeks in the case of respiratory and gastrointestinal illnesses.[35] According to a study published in a medical journal, it took weeks for official sources in Haiti to report details of a cholera epidemic in 2010, resulting in more than 7,000 casualties and 500,000 infections, whereas on Twitter, news of the disease traveled far more quickly.[36]

---

[31] Rita Rubin, How did Vioxx debacle happen?, USA TODAY, October 12, 2004, http://www.usatoday.com/news/health/2004-10-12-vioxx-cover_x.htm.

[32] *See* short description in Andrew Kanter, David Spencer; Malcolm Steinberg, Robert Soltysik, Paul Yarnold & Neil Graham, Supplemental Vitamin B and Progression to AIDS and Death in Black South African Patients Infected With HIV, 21(3) JOURNAL OF ACQUIRED IMMUNE DEFICIENCY SYNDROMES 252 (1999).

[33] *Ibid*.

[34] Jeremy Ginsberg, Matthew Mohebbi, Rajan Patel, Lynnette Brammer, Mark Smolinski & Larry Brilliant, Detecting Influenza Epidemics Using Search Engine Query Data, 457 NATURE 1012 (2009).

[35] Brian Fung, Using Data Mining to Predict Epidemics Before They Spread, THE ATLANTIC, May 2, 2012, http://www.theatlantic.com/health/archive/2012/05/using-data-mining-to-predict-epidemics-before-they-spread/256605.

[36] Rumi Chunara, Jason Andrews & John Brownstein, Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak, 86 AM. J. TROP. MED. HYG. 39 (2012); *also see* Alessio Signorini, Alberto Maria Segre & Philip Polgreen, The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza a H1N1 Pandemic, PLoS ONE, May 2011, http://www.divms.uiowa.edu/~asignori/papers/use-twitter-track-level-disease-activity-and-concern-in-us-during-h1n1.pdf.

With all this in mind, English Prime Minister David Cameron recently announced that every NHS patient would henceforth be a "research patient" whose medical record will be "opened up" for research by private healthcare firms.[37] The Prime Minister emphasized that privacy-conscious patients would be given opt out rights. He added that "this does not threaten privacy, it doesn't mean anyone can look at your health records, but it does mean using anonymous data to make new medical breakthroughs."

Mobile

While a significant driver for research and innovation, the health sector is not the only arena for groundbreaking big data use. A group of scientists working on a collaborative project at MIT, Harvard and additional research universities is currently analyzing mobile phone communications to better understand the needs of the 1 billion people who live in settlements or slums in developing countries.[38] They explore ways to predict food shortages using variables such as market prices, drought, migrations, previous regional production, and seasonal variations;[39] to quantify crime waves, by tracking the time, place, and nature of criminal activity in locations across a city; [40] and to decide which intervention works best to improve learning outcomes in developing country schools.[41]

Smart grid

Another example is the "smart grid", which refers to the modernization of the current electrical grid to introduce a bi-directional flow of information and electricity.[42] The smart grid is designed to allow electricity service providers, users, and other third parties to monitor and control electricity use. Utilities view the smart grid as a way to precisely locate power outages or other problems, including cyber-attacks or natural disasters, so that technicians can be dispatched to mitigate problems. Consumers benefit from more choices on how, when, and how much electricity they use.[43] Pro-environment policymakers view the smart grid as key to providing better power quality, and more efficient delivery of electricity to facilitate the move toward

---

[37] Everyone 'to be research patient', says David Cameron, BBC, December 5, 2011, http://www.bbc.co.uk/news/uk-16026827.

[38] Big Data for Social Good Initiative, Engineering Social Systems, http://www.hsph.harvard.edu/ess/bigdata.html; *see, e.g.,* Amy Wesolowski & Nathan Eagle, Parameterizing the Dynamics of Slums, AAAI Spring Symposium 2010 on Artificial Intelligence for Development (AI-D), http://ai-d.org/pdfs/Wesolowski.pdf.

[39] Washington Okori & Joseph Obua, Machine Learning Classification Technique for Famine Prediction, Proceedings of the World Congress on Engineering 2011, http://www.iaeng.org/publication/WCE2011/WCE2011_pp991-996.pdf.

[40] Jameson Toole, Nathan Eagle and Joshua Plotkin, Quantifying Crime Waves, AAAI Spring Symposium 2010 on Artificial Intelligence for Development (AI-D), http://ai-d.org/pdfs/Toole.pdf.

[41] Massoud Moussavi and Noel McGinn, A Model for Quality of Schooling, American Associations for the Advancement of Artificial Intelligence, 2010 AAAI Spring Symposium Series.

[42] Information and Privacy Commissioner of Ontario & Future of Privacy Forum, Smart Privacy for the Smart Grid: Embedding Privacy into the Design of Electricity Conservation, November 2009, http://www.ipc.on.ca/images/Resources/pbd-smartpriv-smartgrid.pdf.

[43] Katie Fehrenbacher, Introducing the Facebook Social Energy App, October 17, 2011, GIGAOM http://gigaom.com/cleantech/introducing-the-facebook-social-energy-app.

renewable energy. Other benefits, such as accurately predicting energy demands to optimize renewable sources are reaped by society at large.

Traffic management

An adjacent area for data driven environmental innovation is traffic management and control. Governments around the world are establishing electronic toll pricing systems, which determine differentiated payments based on mobility and congestion charges.[44] Users pay depending on their use of vehicles and roads. Urban planners benefit from the analysis of personal location data for decisions involving road and mass-transit construction, mitigation of traffic congestion, and planning for high-density development.[45] Such decisions can not only cut congestion but also control the emission of pollutants.[46] At the same time, individual drivers benefit from smart routing based on real-time traffic information, including accident reports and information about scheduled roadwork and congested areas. Automotive telematics is another area of innovation. Vehicles are equipped with navigation systems with embedded communication modules proposing a range of telematics services to offer drivers fuel-efficient driving, planning trips taking into account the location of charging stations, or activating their air conditioner remotely.[47]

Retail

Big data is also transforming the retail market. It was Wal-Mart's inventory-management system, called Retail Link, which pioneered the age of big data by enabling suppliers to see the exact number of their products on every shelf of every store at each precise moment in time.[48] Many of us use Amazon's "Customers Who Bought This Also Bought" feature, prompting users to consider buying additional items selected by a collaborative filtering tool. The fundamental business model of the Internet is financing products and services with targeted ads whose value correlates directly with the amount of information collected from users.[49] Businesses care not so much about the identity of each individual user but rather on the attributes of her profile, which

---

[44] In Europe: Directive 2004/52/EC of the European Parliament and of the Council of 29 April 2004 on the interoperability of electronic road toll systems in the Community; Commission Decision 2009/750/EC of 6 October 2009 on the definition of the European Electronic Toll Service and its technical element.

[45] *See, e.g.*, Carlo Ratti, Riccardo Maria Pulselli, Sarah Williams & Dennis Frenchman, Mobile Landscapes: Using Location Data from Cell-Phones for Urban Analysis, 33 ENVIRONMENT AND PLANNING B: PLANNING AND DESIGN 727 (2006).

[46] MGI Report, *supra* note 21.

[47] Sastry Duri, Jeffrey Elliot, Marco Gruteser, Xuan Liu, Paul Moskowitz, Ronald Perez, Moninder Singh & Jung-Mu Tang, Data Protection and Data Sharing in Telematics, 9(6) MOBILE NETWORKS & APPLICATIONS, 693 (2004).

[48] A different game: Information is transforming traditional businesses, THE ECONOMIST, February 25, 2010, http://www.economist.com/node/15557465.

[49] *See* Article 29 Working Party, Opinion 2/2010 on online behavioral advertising (WP 171), 22 June 2010, http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2010/wp171_en.pdf; FTC Staff Report, Self-Regulatory Principles for Online Behavioral Advertising, February 2009, http://www.ftc.gov/os/2009/02/P085400behavadreport.pdf.

determine the nature of ads she is shown.[50] Analytics can also be used in the offline environment to study customers' in-store behavior to improve store layout, product mix, and shelf positioning. McKinsey reports that "[r]ecent innovations have enabled retailers to track customers' shopping patterns (*e.g.*, footpath and time spent in different parts of a store), drawing real-time location data from smartphone applications (*e.g.*, Shopkick), shopping cart transponders, or passively monitoring the location of mobile phones within a retail environment."[51] Increasingly, organizations are seeking to link online activity to offline behavior, both in order to assess the effectiveness of online ad campaigns as judged by conversion to in-store purchases and to re-target in-store customers with ads when they go online.

Payments

Another major arena for valuable big data use is fraud detection in the payment card industry. With electronic commerce capturing an increasingly large portion of the retail market, merchants, who bear ultimate responsibility for fraudulent card payments,[52] must implement robust mechanisms to identify suspect transactions often performed by first time customers. To this end, companies developed solutions to provide merchants with predictive fraud scores for "Card-Not-Present transactions" in order to measure in real time the likelihood that a transaction is fraudulent. To do that, the services analyze buyer histories and provide evaluations, much like a summarized list of references but in the form of a single score. As fraudsters become more sophisticated in their approach, online merchants must remain ever more vigilant in their efforts to protect the integrity of the online shopping experience.

Online

Finally, consider the massive data silos maintained by the online tech giants, Google, Facebook, Microsoft, Apple and Amazon. These companies amass previously unimaginable amounts of personal data. Facebook, for example, has more than 900 million users who upload more than

---

[50] Omer Tene & Jules Polonetsky, To Track or 'Do Not Track': Advancing Transparency and Individual Control in Online Behavioral Advertising, 13 MINN. J. L. SCI. & TECH. 282 (2012).

[51] MGI Report, *supra* note 21, at p. 68.

[52] A set of laws and regulations serve to protect consumer users of credit and debit cards from bearing the consequences of fraud losses associated with lost or stolen cards. See the Truth in Lending Act (TILA), which is contained in Title I of the Consumer Credit Protection Act (15 U.S.C. § 1601 et seq.) together with Regulation Z, promulgated by the Federal Reserve Board pursuant to authority granted under 15 U.S.C. § 1607; as well as the Electronic Fund Transfer Act (EFTA) (15 U.S.C. § 1693 et seq.), together with Federal Reserve Board Regulation E. The EFTA and Regulation E place a floating cap on a consumer cardholder's liability for unauthorized debit card use under which the maximum liability amount is determined when the cardholder notifies the card issuer of the loss or theft of the card used to perpetrate the fraud. If the cardholder notifies the card issuer within two business days of learning of the loss or theft of the debit card, the cardholder's maximum liability is limited to the lesser of the actual amount of unauthorized transfers or $50. 12 C.F.R. § 205.6(b)(1). Liability is further allocated between card issuers and merchants, generally shifting the risk away from the card issuers and onto the merchants, based on a complicated set of rules that vary based on the type of transaction at issue. See Duncan Douglass, An Examination of the Fraud Liability Shift in Consumer Card-Based Payment Systems, FEDERAL RESERVE BANK OF CHICAGO (2009), www.chicagofed.org/digital.../ep_1qtr2009_part7_douglass.pdf.

250 millions photos and click the "like" button more than 2.5 billion times per day.[53] Google offers a plethora of data-intensive products and services, including its ubiquitous search engine, Android operating system, Chrome browser, Gmail, Youtube, Google Maps, Google Plus, Google Analytics, Google Apps, and many others.[54] In addition, Google owns the largest online ad serving company DoubleClick, which it purchased in 2007, much to the consternation of privacy advocates,[55] as well as AdMob, the leading mobile advertising company. As a result, Google now has a presence on well over 70 percent of third party websites.[56] Amazon and Yahoo are seeking new ways to leverage and monetize their treasure trove of customer data.[57] Apple and Microsoft make operating systems as well as browsers, both of which are important focal points for collecting online and mobile user information.

Big Data: Big Concerns

Big data poses big privacy risks. The harvesting of large sets of personal data and use of state of the art analytics clearly implicate growing privacy concerns. Protecting privacy become harder as information is multiplied and shared ever more widely among multiple parties around the world. As more information regarding individuals' health, financials, location, electricity use and online activity percolates, concerns arise about profiling, tracking, discrimination, exclusion, government surveillance and loss of control.[58] This Part lays out some of the unique privacy risks presented by big data.

Incremental effect

The accumulation of personal data has an incremental adverse effect on privacy.[59] A researcher will draw entirely different conclusions from a string of online search queries consisting of the words "paris", "hilton" and "louvre" as compared to one featuring "paris", "hilton" and "nicky". Add thousands and thousands of search queries, and you can immediately sense how the data

---

[53] Melissa Fach, Stats on Facebook 2012, Search Engine Journal, February 17, 2012, http://www.searchenginejournal.com/stats-on-facebook-2012-infographic/40301; also see 10 Key Statistics About Facebook, EXPERIAN HITWISE BLOG, February 2, 2012, http://www.experian.com/blogs/hitwise/2012/2/2/10-key-statistics-about-facebook.

[54] http://www.google.com/intl/en/about/products/index.html.

[55] *See* Statement of the Federal Trade Commission Concerning Google/DoubleClick, FTC File No. 071-0170, December 20, 2007, http://www.ftc.gov/os/caselist/0710170/071220statement.pdf; *and see* Dissenting Statement of Commissioner Pamela Jones Harbour, http://www.ftc.gov/os/caselist/0710170/071220harbour.pdf.

[56] Balachander Krishnamurthy & Craig Wills, Privacy Diffusion on the Web: A Longitudinal Perspective, October 2009, submitted as public comment to Federal Trade Commission Exploring Privacy Roundtable Series, http://www.ftc.gov/os/comments/privacyroundtable/544506-00009.pdf.

[57] Nicole Perlroth, Revamping at Yahoo to Focus on Its Media Properties and Customer Data, NY TIMES, April 11, 2012, http://bit.ly/HUneMM.

[58] For a taxonomy of privacy harms, *see* Daniel Solove, A Taxonomy of Privacy, 154 U. PA. L. REV. 477 (2006).

[59] Solove in his "taxonomy" calls this "aggregation". *Ibid*.

become ever more revealing.[60] Moreover, once data – such as a clickstream or a cookie number – are linked to an identified individual, they become difficult to disentangle.[61] This was demonstrated by University of Texas researchers Arvind Narayanan and Vitaly Shmatikov, who re-associated de-identified Netflix movie recommendations with identified individuals by crossing a de-identified database with publicly available resources available online.[62] Narayanan and Shmatikov explained: "once any piece of data has been linked to a person's real identity, any association between this data and a virtual identity breaks the anonymity of the latter."[63] Paul Ohm warmed that this incremental effect will lead to a "database of ruin", chewing away, bit by byte, on an individual's privacy until his or her profile is completely exposed.[64]

More generally, the effervescent nature of personal data makes it difficult to recapture after it is exposed in the public or semi-public sphere.[65] This is why the European Commission's proposal of a "right to be forgotten", which would allow individuals to demand organizations to wipe their data slate clean,[66] has been met with fierce resistance from online platforms[67] and free speech advocates,[68] who are concerned about its effect on the delicate balance between privacy and regulation on the Internet.

---

[60] *See, e.g.*, Michael Barbaro & Tom Zeller, A Face Is Exposed for AOL Searcher No. 4417749, NY TIMES, August 9, 2006, http://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=all.

[61] *See* Agreement Containing Consent Order, File No. 102-3058, In the Matter of Myspace, Federal Trade Commission, May 8, 2012, http://www.ftc.gov/os/caselist/1023058/120508myspaceorder.pdf (charging Myspace "constructively shared" personally identifiable information with third party advertisers by sharing with such advertisers a unique identifier assigned to the profile of each Myspace user (a "Friend ID"), which could then be used to access such user's profile information – a practice referred to in the industry as "cookie syncing"). For an analysis of "cookie syncing" *see* Ed Felten, Syncing and the FTC's Myspace Settlement, TECH@FTC, May 8, 2012, http://techatftc.wordpress.com/2012/05/08/syncing-and-the-ftcs-myspace-settlement.

[62] Arvind Narayanan & Vitaly Shmatikov, Robust De-anonymization of Large Sparse Datasets, 2008 IEEE SYMPOSIUM ON SECURITY & PRIVACY 111

[63] *Ibid*, at p. 119.

[64] Paul Ohm, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, 57 UCLA L. REV. 1701 (2010).

[65] A social networking service (SNS) is a semi-public sphere. While an individual user's postings are made according to the SNS privacy settings, other users are not subject to a legal obligation to comply with such user's individual settings. Consequently, a posting made by a user and restricted to her "friends" may later be disseminated broadly by those friends so as to become public or semi-public. *See* Omer Tene, Me, Myself and I: Aggregated and Disaggregated Identities on Social Networking Services, __ J. INT'L COMM. L. & TECH. (forthcoming, 2012).

[66] European Data Protection Regulation, art. 17; *also see* Viviane Reding, Vice President, European Commission, The EU Data Protection Reform 2012: Making Europe the Standard Setter for Modern Data Protection Rules in the Digital Age 5, January 22, 2012, http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/12/26&format=PDF.

[67] Peter Fleischer, Foggy Thinking About the Right to Oblivion, PRIVACY . . . ? BLOG, March 9, 2011, http://peterfleischer.blogspot.com/2011/03/foggy-thinking-about-right-to-oblivion.html.

[68] *See* Jeffrey Rosen, The Right to Be Forgotten, 64 STAN. L. REV. ONLINE 88 (2012).

Automated decision-making

The relegation of decisions about an individual's life to automated processes based on algorithms and artificial intelligence raises concerns about discrimination, self-determination, and the narrowing of choice.[69] This is true not only for decisions relating to an individual's credit, insurance or job prospects, which for many years have been regulated by laws such as the Fair Credit Reporting Act,[70] but also for highly customized choices regarding which advertisements or content a user will see.[71] In his book "*The Daily You: How The New Advertising Industry Is Defining Your Identity And Your Worth*", Joseph Turow argues that increased personalization based on opaque corporate profiling algorithms poses a risk to open society and democratic speech.[72] He explains that by "pigeonholing" individuals into pre-determined categories, automated decision making compartmentalizes society into pockets (or "echo chambers") of likeminded individuals.[73] Turow believes that government should regulate information intermediaries to ensure that users have full control over their data and content consumption.

Predictive analysis

Big data may facilitate predictive analysis with stark implications for individuals susceptible to disease, crime, or other socially stigmatizing characteristics or behaviors. To be sure, predictive analysis can be used for societally beneficial goals, such as planning disaster recovery in an earthquake prone area based on individuals' evacuation paths and purchase needs. Yet it can easily cross the "creepiness" threshold.

Consider a recent story in the *New York Times*, which uncovered that retailing giant Target assigns a "pregnancy prediction score" to customers based on their purchase habits.[74] According to the *Times,* Target employed statisticians to sift back through historical buying records of women who had signed up for baby registries. The statisticians discovered latent patterns, such as women's preference for unscented lotion around the beginning of their second trimester or a tendency to buy supplements like calcium, magnesium and zinc within the first 20 weeks of a pregnancy. They were able to determine a set of products that, when grouped together, allowed them to extremely accurately predict a customer's pregnancy and due date. In one case, the *Times* reported, a father of a teenage girl stormed into a Target store to complain that his daughter received coupons and advertisements for baby products. A few days later, he called

---

[69] *See* Ruth Gavison, Privacy, 89 YALE L.J. 421 (1980); European Data Protection Directive, art. 15.

[70] 15 U.S.C. § 1681 et seq.

[71] Kashmir Hill, Resisting The Algorithms, FORBES, May 5, 2011, http://www.forbes.com/sites/kashmirhill/2011/05/05/resisting-the-algorithms.

[72] Joseph Turow, THE DAILY YOU: HOW THE NEW ADVERTISING INDUSTRY IS DEFINING YOUR IDENTITY AND YOUR WORTH (Yale University Press, 2011). For similar arguments *also see* Eli Pariser, THE FILTER BUBBLE: WHAT THE INTERNET IS HIDING FROM YOU (Penguin, 2011). .

[73] This phenomenon, sometimes referred to as "cyberbalkanization" (*see* Wikipedia entry, http://en.wikipedia.org/wiki/Cyberbalkanization), was originally explored by Cass Sunstein, REPUBLIC.COM (Princeton University Press, 2001); *also see* Andrew Shapiro, THE CONTROL REVOLUTION: HOW THE INTERNET IS PUTTING INDIVIDUALS IN CHARGE AND CHANGING THE WORLD WE KNOW (PublicAffairs, 2000).

[74] Charles Duhigg, How Companies Learn Your Secrets, NY TIMES MAGAZINE, February 16, 2012, http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all.

the store manager to apologize, admitting that: "there's been some activities in my house I haven't been completely aware of. She's due in August."[75]

Predictive analysis is useful for law enforcement, national security, credit screening, insurance and employment. It raises ethical dilemmas captured, for example, in the film *Minority Report*, where a "PreCrime" police department apprehends "criminals" based on foreknowledge of their future misdeeds. It could facilitate unlawful activity such as "redlining" – denying or increasing the cost of services such as loans, insurance, or healthcare to residents of neighborhoods comprised mostly of minorities. Although these practices are illegal under current laws, critics expressed concerns that data is surreptitiously being used in such a manner.[76]

Predictive analysis is particularly problematic when based on sensitive categories of data, such as health, race, or sexuality. It is one thing to recommend for a customer books, music or movies she might be interested in based on her previous purchases;[77] it is quite another thing to identify when she is pregnant before her closest family knows. In the law enforcement arena, it raises the specter of surveying or even incarcerating individuals based on thoughts as opposed to deeds.[78] This type of activity, while clearly unconstitutional under existing United States law, is not so far-fetched in other parts of the world,[79] and could conceivably cross the line from fiction to reality given the right circumstances in the United States.[80]

Even with non-sensitive data categories, predictive analysis may have a stifling effect on individuals and society, perpetuating old prejudices. The rich and well educated will get the fast track; the poor and underprivileged will have the deck stacked against them even more so than before. By ignoring outliers and assuming that "what has been is what will be",[81] predictive

---

[75] *Ibid*.

[76] *See, e.g.*, Joint Filing to the Federal Trade Commission of Center for Digital Democracy, U.S. PIRG & World Privacy Forum, In the Matter of Real-time Targeting and Auctioning, Data Profiling Optimization, and Economic Loss to Consumers and Privacy, April 8, 2010, http://www.centerfordigitaldemocracy.org/sites/default/files/20100407-FTCfiling.pdf.

[77] Consider Amazon, Netflix and Pandora recommendation systems; Gediminas Adomavicius & Alexander Tuzhilin, Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, 17(6) IEEE Transactions on Knowledge & Data Engineering (June 2005); *but see* Ryan Singel, Netflix Spilled Your Brokeback Mountain Secret, Lawsuit Claims, Wired, December 17, 2009, http://www.wired.com/threatlevel/2009/12/netflix-privacy-lawsuit (plaintiff argues that Netflix made it possible for her to be "outed" when it disclosed insufficiently anonymous information about her viewing habits, including films from the "Gay & Lesbian" genre).

[78] Rosamunde van Brakel & Paul De Hert, Policing, surveillance and law in a pre-crime society: Understanding the consequences of technology based strategies, 20 J. POLICE STUDIES 163 (2011).

[79] *See, e.g.*, Clive Thompson, Google's China Problem (and China's Google Problem), NY TIMES MAGAZINE, April 23, 2006, http://www.nytimes.com/2006/04/23/magazine/23google.html?pagewanted=all. Google has since withdrawn from the Chinese market. David Drummond, A new approach to China, GOOGLE OFFICIAL BLOG, January 12, 2010.

[80] Marc Rotenberg, Foreword: Privacy and Secrecy after September 11, 86 MINN. L. REV. 1115 (2002).

[81] ECCLESIASTES 1:9 (New Revised Standard Version).

analysis becomes a self-fulfilling prophecy that accentuates social stratification.[82] It leads to morally contentious conclusions, such as those drawn by the (in)famous 2001 article of John Donohue and Steven Levitt, "*The Impact of Legalized Abortion on Crime*", which argued that the legalization of abortion in the 1970s contributed significantly to reductions in crime rates experienced in the 1990s.[83]

Lack of Access and Exclusion

An additional concern raised by big data is that it tilts an already uneven scale in favor of organizations and against individuals. The big benefits of big data, the argument goes, accrue to government and big business, not to individuals – and they often come at individuals' expense. In the words of the adage, "if you're not paying for it, you're not the customer; you're the product".

The exclusion of individuals from the benefits of the use of their data is manifest in two main ways. First, online interactions are barter-like transactions where individuals exchange personal data for free services.[84] Yet those transactions appear to take place in an inefficient market hampered by steep information asymmetries, which are further aggravated by big data. Transacting with a big data platform is like a game of poker where one of the players has his hand open and the other keeps his cards close. The online company knows the preferences of the transacting individual inside out, perhaps better than the individual knows him or herself. It can therefore usurp the entire value surplus available in the transaction by pricing goods or services as close as possible to the individual's reservation price.

Second, organizations are seldom prepared to share the wealth created by individuals' personal data with those individuals. Sir Tim Berners-Lee recently remarked:

> "My computer has a great understanding of my state of fitness, of the things I'm eating, of the places I'm at. My phone understands from being in my pocket how much exercise I've been getting and how many stairs I've been walking up and so on. Exploiting such data could provide hugely useful services to individuals, he said, but only if their computers had access to personal data held about them by web companies. One of the issues of social networking silos is that they have the data and I don't."[85]

---

[82] Jay Stanley, Eight Problems With "Big Data", ACLU BLOG, April 25, 2012, https://www.aclu.org/blog/technology-and-liberty/eight-problems-big-data.

[83] John Donohue & Steven Levitt, The Impact of Legalized Abortion on Crime, 66 QUARTERLY J. ECON. 379 (2001); for criticism *see, e.g.*, Christopher Foote & Christopher Goetz, The Impact of Legalized Abortion on Crime: Comment, Federal Reserve Bank of Boston Working Paper 05-15 (November 2005), http://www.bos.frb.org/economic/wp/wp2005/wp0515.pdf (distinguishing the role of abortion from other potential influences on crime, some of which vary year by year or state by state, including for example the "crack" epidemic, which rose and receded at different times in different places).

[84] Chris Anderson, FREE: THE FUTURE OF A RADICAL PRICE (Hyperion Books, 2009).

[85] Ian Katz, Tim Berners-Lee: Demand your Data from Google and Facebook, THE GUARDIAN, April 18, 2012, http://www.guardian.co.uk/technology/2012/apr/18/tim-berners-lee-google-facebook. *Also see* Bruce Upbin, How Intuit Uses Big Data For The Little Guy, FORBES, April 26, 2012,

The right of access granted to individuals under the European Data Protection Directive[86] and additional fair information principles has been implemented narrowly. Even where they comply with the law, organizations provide individuals with little useful information.

The Ethics of Analytics: Drawing the Line

Like any other type of research, data analytics can cross the threshold of unethical behavior. Consider the recent research by a Texas University developmental psychology professor, who logged and reviewed every text message, email, photo, and instant message sent by a group of 175 teenagers on Blackberries that she provided to them.[87] While the participants and their parents were required to sign consent forms, it is doubtful that the minors were sufficiently informed to assess the full implications of the omniscient surveillance.[88] Like children's data, other categories of sensitive data may be collected and analyzed for ethically dubious research. Consider a service analyzing individuals' preferences on pornography sites for use in behavioral advertising.[89] More complicated yet, the analysis of apparently innocuous data may create new sensitive facts about an individual, such as Target's "pregnancy score"[90] or a prediction of the onset of Alzheimer's. Where should the red lines be drawn when it comes to big data analysis? Moreover, who should benefit from access to big data? Could ethical scientific research be conducted without disclosing to the general public the data used to reach the results?

Chilling Effect

As recently observed by Jay Stanley of the ACLU, "as the ramifications of big data analytics sink in, people will likely become much more conscious of the ways they're being tracked, and the chilling effects on all sorts of behaviors could become considerable."[91] The result is what the

---

http://www.forbes.com/sites/bruceupbin/2012/04/26/how-intuit-uses-big-data-for-the-little-guy, writing: "Big Data means big challenges and big opportunities. But, hey, what about me? What do I (meaning the average joe) get out of all this? Companies are flying on the contrails of our spending, hiring and networking behavior, especially at the social/mobile colossi like Facebook, Google and Apple. We ought to see some of that value. Rather than just take take take, why can't more companies give back, reflect our data back on us? Doing this in a real, honest way has to create some business value".

[86] European Data Protection Directive, art. 12; White House Blueprint, Consumer Privacy Bill of Rights principle of "Access and Accuracy".

[87] Kashmir Hill, A Texas University's Mind-Boggling Database Of Teens' Daily Text Messages, Emails, and IMs Over Four Years, FORBES, April 18, 2012, http://www.forbes.com/sites/kashmirhill/2012/04/18/a-texas-universitys-mind-boggling-database-of-teens-daily-text-messages-emails-and-ims-over-four-years/.

[88] See, e.g., Michael Zimmer, Research Ethics and the Blackberry Project, MichaelZimmer.org, April 25, 2012, http://michaelzimmer.org/2012/04/25/research-ethics-and-the-blackberry-project.

[89] Kashmir Hill, History Sniffing: How YouPorn Checks What Other Porn Sites You've Visited and Ad Networks Test The Quality of Their Data, FORBES, November 30, 2010, http://www.forbes.com/sites/kashmirhill/2010/11/30/history-sniffing-how-youporn-checks-what-other-porn-sites-youve-visited-and-ad-networks-test-the-quality-of-their-data.

[90] *Supra* note 74 and accompanying text.

[91] Jay Stanley, The Potential Chilling Effects of Big Data, ACLU BLOG, April 30, 2012, https://www.aclu.org/blog/technology-and-liberty/potential-chilling-effects-big-data.

former UK privacy regulator dubbed "a surveillance society", a psychologically oppressive world in which individuals are cowed to conforming behavior by the state's potential panoptic gaze.[92]

The Legal Framework: Challenges

How does the existing privacy framework deal with the big data phenomenon? This Part reviews the FIPPs strained by the current technological and business landscape, including the definition of PII; the principles of data minimization and purpose limitation; and the concept of consent.[93] It argues that inevitably, these elements of the privacy framework should adjust to reflect existing technological and organizational realities, which include ubiquitous data collection and individuals who are ill placed to meaningfully review privacy policies. Together with the next Part, it argues that the FIPPs should be used as a set of levers, which can be modulated to address big data by relaxing the principles of data minimization and individual control while tightening requirements for transparency, access and accuracy.

Definition of PII

Traditionally, de-identification was viewed as a silver bullet allowing organizations to reap the benefits of analytics while preserving individuals' privacy.[94] Organizations used various methods of de-identification (anonymization, pseudonymization, encryption, key-coding, data sharding) to distance data from real identities.[95] Yet, over the past few years, computer scientists have repeatedly shown that even anonymized data can typically be re-identified and associated with specific individuals.[96] De-identified data, in other words, is a temporary state rather than a

---

[92] Watchdog's Big Brother UK warning, BBC, August 16, 2004, http://news.bbc.co.uk/2/hi/uk_news/politics/3568468.stm, quoting UK Information Commissioner Richard Thomas as saying the UK could "sleepwalk into a surveillance society". The paradigmatic example is, of course, George Orwell, 1984 (1948).

[93] This article deals with the adjustment of the existing privacy framework to accommodate big data realities. Other privacy issues raised by big data, such as government access to or surveillance of private sector databases, are beyond the scope of this paper. *See, e.g.*, James Dempsey & Lara Flint, Commercial Data and National Security, 72 GEO. WASH. L. REV. 1459 (2004).

[94] *See, e.g.,* Article 29 Data Protection Working Party, Opinion 4/2007 on the Concept of Personal Data, WP 136, June 20, 2007, http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf; Markle Foundation Task Force on National Security in the Information Age, Creating a Trusted Network for Homeland Security (2003), http://www.markle.org/downloadable_assets/nstf_report2_full_report.pdf; Ira Rubinstein, Ronald Lee & Paul Schwartz, Data Mining and Internet Profiling: Emerging Regulatory and Technological Approaches, 75 U. CHI. L. REV. 261, 268-29 (2008).

[95] W. Kuan Hon, Christopher Millard & Ian Walden, The Problem of 'Personal Data' in Cloud Computing - What Information is Regulated? The Cloud of Unknowing, Part 1, March 15, 2011, Queen Mary School of Law Legal Studies Research Paper No. 75/2011, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1783577.

[96] This line of research was pioneered by Latanya Sweeney and made accessible to lawyers by Paul Ohm. Ohm, *supra* note 64; Latanya Sweeney, Uniqueness of Simple Demographics in the U.S. Population, Laboratory for International Data Privacy Working Paper, LIDAP-WP4 (2000); Narayanan & Shmatikov,

stable category.[97] In an influential law review article, Paul Ohm observed that "[r]e-identification science disrupts the privacy policy landscape by undermining the faith that we have placed in anonymization."[98] The implications for government and businesses can be stark, since de-identification has become a key component of numerous business models, most notably in the context of health data (*e.g.*, clinical trials), online behavioral advertising, and cloud computing.

The first major policy question raised by the big data phenomenon concerns the scope of information subject to privacy law. How robust must de-identification be in order to "liberate" data from the throes of privacy legislation? One possible conclusion, apparently supported by Ohm himself, is that all data should be treated as PII and subjected to the regulatory framework.[99] Yet such a result would create perverse incentives for organizations to forgo de-identification altogether and therefore increase, not alleviate, privacy and data security risks.[100] A further pitfall is that with a vastly expanded definition of PII, the privacy framework would become all but unworkable. Difficult enough to comply with and enforce today, the current framework may well be unmanageable if it extends to every piece of information.[101] Moreover, while anonymized information always carries some risk of re-identification, many of the most pressing privacy risks exist only if there is reasonable likelihood of re-identification. As uncertainty is introduced into the re-identification equation, we cannot know whether the information truly corresponds to a particular individual, and the dataset becomes more anonymous as larger amounts of uncertainty are introduced.[102]

---

supra note 62; Arvind Narayanan & Vitaly Shmatikov, Myths and Fallacies of "Personally Identifiable Information", 53(6) COMMUNICATIONS OF THE ACM 24 (2010), http://www.cs.utexas.edu/users/shmat/shmat_cacm10.pdf; and more recently Arvind Narayanan, Hristo Paskov *et al*, On the Feasibility of Internet-Scale Author Identification, IEEE S&P 2012, http://randomwalker.info/publications/author-identification-draft.pdf.

[97] *See* series of blog posts by Ed Felten, the FTC's Chief Technologist and a Professor of Computer Science at Princeton: Ed Felten, Does Hashing Make Data "Anonymous"?, TECH@FTC, April 22, 2012, http://techatftc.wordpress.com/2012/04/22/does-hashing-make-data-anonymous; Ed Felten, Are pseudonyms "anonymous"?, TECH@FTC, April 30, 2012, http://techatftc.wordpress.com/2012/04/30/are-pseudonyms-anonymous; Felten, *supra* note 61.

[98] Ohm, *supra* note 64, at 1704.

[99] *Ibid*, at 1742.

[100] Ann Cavoukian & Khaled El Emam, Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy, Information & Privacy Commissioner of Ontario Whitepaper, June 2011, http://www.ipc.on.ca/images/Resources/anonymization.pdf.

[101] For example, according to a 2010 report by the EU Agency for Fundamental Rights, even in Europe, data protection authorities lack sufficient independence and funding; impose few sanctions for violations of data protection laws; and "are often not equipped with full powers of investigation and intervention or the capacity to give legal advice or engage in legal proceedings." European Union Agency for Fundamental Rights, Data Protection in the European Union: The role of National Data Protection Authorities, May 7, 2010, http://fra.europa.eu/fraWebsite/attachments/Data-protection_en.pdf.

[102] Betsy Masiello & Alma Whitten, Engineering Privacy in an Age of Information Abundance, 2010 AAAI Spring Symposium Series, 2010, http://www.aaai.org/ocs/index.php/SSS/SSS10/paper/viewFile/1188/1497.

More important, if information, ostensibly not about individuals, comes under full remit of privacy laws based on a remote possibility of it being linked to an individual at some point in time through some conceivable method, no matter how unlikely to be used, many beneficial uses of data would be severely curtailed.[103] Such an approach presumes a value judgment has been made in favor of individual control over highly beneficial uses of data, such as Dr. Altman's discovery of the Paxil-Pravachol side effect; yet it is doubtful that such a value choice has consciously been made.

PII should instead be defined based on a risk matrix taking into account the risk, intent and potential consequences of re-identification, as opposed to a dichotomy between "identifiable" and "non-identifiable" data.[104] A bi-polar approach based on labeling information either "personally identifiable" or not is unhelpful and inevitably leads to an inefficient arms race between de-identifiers and re-identifiers. In this process, the integrity, accuracy and value of the data may be degraded or lost, together with some of its potential societal benefits.[105]

We suggest, first, that the identifiability of data should be viewed as a continuum as opposed to the current dichotomy. This means adopting a scaled approach, under which data that are only identifiable at great cost would remain within the legal framework yet subject to only a subset of fair information principles.[106] Second, we support the approach proposed by the Federal Trade Commission (FTC) in its recent report *Protecting Consumer Privacy in an Era of Rapid Change*,[107] which overlays the statistical probability of re-identifiablity with legally enforceable organizational commitments as well as downstream contractual obligations not to re-identify or to attempt to re-identify. According to the FTC:

> "[A]s long as (1) a given data set is not reasonably identifiable, (2) the company publicly commits not to re-identify it, and (3) the company requires any downstream users of the data to keep it in de-identified form, that data will fall outside the scope of the framework."[108]

Recognizing that it is virtually impossible to guarantee privacy by scrutinizing the data alone, without defining and analyzing its intended uses, the FTC shifts the crux of the inquiry from a

---

[103] *See, e.g.*, Kathleen Benitez & Bradley Malin, Evaluating Re-identification Risks with respect to the HIPAA Privacy Rule, 17 J. AM. MED. INFORM. ASSOC. 169 (2010) (demonstrating actual risk of re-identification may be low).

[104] Paul Schwartz & Daniel Solove, The PII Problem: Privacy and a New Concept of Personally Identifiable Information, 86 NYU L. REV. 1814 (2011); Omer Tene, The Complexities of Defining Personal Data: Anonymization 8(8) DATA PROTECTION L. & POLICY 6 (2011).

[105] Daniel Barth-Jones, Balancing Privacy Protection with Scientific Accuracy: Challenges for De-identification Practice, First Annual CER Symposium: Responding to the National CER Agenda: Evolving Data Sources and Analytics, June 15, 2010, http://www.lewin.com/~/media/lewin/cer/resources/barth-jones_lewin_cer_symposium_6_15_10.pdf.

[106] Schwartz & Solove, *supra* note 104.

[107] FTC Final Report.

[108] *Ibid*, at p. 22.

factual test of identifiablilty to a *legal* examination of an organization's *intent* and *commitment* to prevent re-identification.

Finally, we advocate viewing de-identification as an important protective measure to be taken under the data security and accountability principles, rather than a solution to the big data conundrum.[109] Organizations collecting and harvesting big data would be wise to de-identify data to the extent possible while not compromising their beneficial use. At the same time, the privacy framework will continue to partially apply to de-identified data, since given the incentive to do so researchers can re-link almost any piece of data back to an individual.

Data Minimization

Through various iterations and formulations, data minimization has remained a fundamental principle of privacy law.[110] Organizations are required to limit the collection of personal data to the minimum extent necessary to obtain their legitimate goals. Moreover, they are required to delete data that are no longer used for the purposes for which they were collected and to implement restrictive policies with respect to the retention of personal data in identifiable form. The big data business model is antithetical to data minimization. It incentivizes collection of more data for longer periods of time. It is aimed precisely at those unanticipated secondary uses, the "crown jewels" of big data. After all, who could have anticipated that Bing search queries would be used to unearth harmful drug interactions?[111]

Here too, legal rules collide against technological and business realities. Organizations today collect and retain personal data through multiple channels including the Internet, mobile, biological and industrial sensors, video, e-mail and social networking tools; amass data collected directly from individuals or from third parties; and harvest private, semi-public (*e.g.*, Facebook) or public (*e.g.*, the electoral roll) sources. Data minimization is simply no longer the market norm.

In considering the fate of data minimization, the principles of privacy law must be balanced against additional societal values such as public health, national security and law enforcement, environmental protection and economic efficiency. A coherent framework should be based on a risk matrix weighing the value of data against potential privacy risks. Where prospective data uses are highly beneficial and privacy risks minimal, the legitimacy of processing should be assumed even if individuals decline (or are not asked) to consent. For example, web analytics – the measurement, collection, analysis and reporting of internet data for purposes of understanding and optimizing web usage – creates rich value by ensuring that products and

---

[109] *See* OECD Guidelines, para. 14 ("Accountability Principle"); White House Blueprint, Consumer Privacy Bill of Rights principle of "Accountability"; Article 29 Data Protection Working Party Opinion 3/2010 on the Principle of Accountability, WP 173, July 13, 2010, http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2010/wp173_en.pdf.

[110] *See* OECD Guidelines, para. 7 ("Collection Limitation Principle"), para. 8 ("Data Quality Principle"); European Data Protection Directive, art. 6(b), (c) and (e); White House Blueprint, Consumer Privacy Bill of Rights principles of "Focused Collection" and "Respect for Context".

[111] *Supra* notes 27-30 and accompanying text.

services can be improved to better serve consumers. Privacy risks are minimal, since analytics, if properly implemented, deals with statistical data, typically in de-identified form.[112] Yet requiring online users to opt into analytics would no doubt severely limit its application and use.

This is not to suggest, of course, that we support data being collected "just in case" it becomes useful or that data collected for one purpose be re-purposed at will. Rather that in a big data world, the principle of data minimization should be interpreted differently, requiring organizations to de-identify data when possible; implement reasonable security measures; and limit uses of data to those that are acceptable from not only an individual but also a societal perspective.

Individual Control and Context

Legal frameworks all over the world continue to emphasize consent, or individual control, as a fundamental principle of privacy law. In the United States, "notice and choice" has been the central axis of privacy regulation for more than a decade.[113] In the European Union, consent remains the most commonly used basis to legitimize data processing under Article 7 of the Data Protection Directive.[114] By emphasizing consent, existing privacy frameworks impose significant, sometimes unrealistic obligations on both organizations and individuals. On the one hand, organizations are expected to explain their data processing activities on increasingly small screens and obtain consent from often-uninterested individuals; on the other hand, individuals are expected to read and understand complicated privacy disclosures, which they have little appetite to consume, and express their "informed" consent.[115] This takes place against an increasingly complex backdrop in which data flows are handled through intricate arrangements involving dense networks of platforms and applications, including contractors, subcontractors

---

[112] Much of the criticism of analytics has been driven by careless practices such as the inadvertent leakage of personal data passed from sites to ad networks, misuse of flash cookies, or concerns that data were being used for behavioral advertising. *See* discussion in Paul Schwartz, Data Protection Law and the Ethical Use of Analytics, PRIVACY & SECURITY LAW REPORT, January 2011, http://www.law.berkeley.edu/files/bclt_Schwartz_Data_Protection_Law_and_the_Ethical_Use_of_Analytics.pdf.

[113] A shift away from notice and choice is underway, as reflected in the Whitehouse Blueprint and FTC Final Report; yet under both frameworks, notice and choice remains a central principle.

[114] European Data Protection Directive, art. 7(a), 8(2)(a); Article 29 Data Protection Working Party, Opinion 15/2011 on the Definition of Consent, WP 187, July 13, 2011, http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2011/wp187_en.pdf; the European Data Protection Regulation would significantly tighten the requirements for consent, effectively permitting only explicit consent and thereby presumably narrowing the scope of consent-based processing; European Data Protection Regulation, art. 4(8) (defining "the data subject's consent" as "any freely given specific, informed and explicit indication of his or her wishes…").

[115] Aleecia McDonald & Lorrie Faith Cranor, The Cost of Reading Privacy Policies, I/S: A JOURNAL OF LAW AND POLICY FOR THE INFORMATION SOCIETY, 2008 Privacy Year in Review issue, http://www.aleecia.com/authors-drafts/readingPolicyCost-AV.pdf (finding that to read every privacy policy encountered, an average individual would need to spend approximately 30 days per year); *also see* Alexis Madrigal, Reading the Privacy Policies You Encounter in a Year Would Take 76 Work Days, THE ATLANTIC, March 1, 2012, http://www.theatlantic.com/technology/archive/2012/03/reading-the-privacy-policies-you-encounter-in-a-year-would-take-76-work-days/253851.

and service providers operating globally. Moreover, to be meaningful, consent must be purpose- (or context-) specific. Yet by its very nature, big data analysis seeks surprising correlations and produces results that resist prediction.

The consent model is flawed from an economic perspective. Information asymmetries and well-documented cognitive biases cast a shadow on the authenticity of individuals' privacy choices. For example, Alessandro Acquisti and his colleagues have shown that simply by providing users a *feeling* of control, businesses can encourage the sharing of data regardless of whether or not users actually gained control.[116] Joseph Turow and others have shown that "[w]hen consumers see the term 'privacy policy,' they believe that their personal information will be protected in specific ways; in particular, they assume that a website that advertises a privacy policy will not share their personal information."[117] In reality, however, this is not the case. Privacy policies often serve more as liability disclaimers for businesses than as assurances of privacy for consumers.

At the same time, collective action problems threaten to generate a suboptimal equilibrium where individuals fail to opt-into societally beneficial data processing in the hope of free riding on others' good will. Consider, for example, Internet browser crash reports, which very few users opt-into not so much because of real privacy concerns but rather due to a (misplaced) belief that others will do the job for them. As is often the case in public opinion polling, the precise wording of choice menus presented to individuals has a disproportionate effect on their decisions to opt-in or out. We suspect, for example, that if Bing users were prompted by the search engine to permit analysis of their search logs for the detection of harmful drug interactions most users would decline. Yet when asked in retrospect about the actions of Dr. Altman and his team, most would find them commendable.

Similar freeriding is common in other contexts where the difference between opt-in and opt-out regimes is stark. This is the case, for example, with organ donation rates: In countries where organ donation is opt-in, donation rates tend to be very low compared to countries that are culturally similar but have an opt-out regime. Consider, for example, the donation rates in Sweden 85.9% (opt-out) and Denmark 4.25% (opt-in); or in Austria 99.9% (opt-out) and Germany 12% (opt-in).[118]

An additional problem is that consent-based processing tends to be regressive, since individuals' expectations fall back on existing experiences. For example, if Facebook had not proactively

[116] Laura Brandimarte, Alessandro Acquisti & George Loewenstein, Misplaced Confidences: Privacy and the Control Paradox, Ninth Annual Workshop on the Economics of Information Security (WEIS) (2010), http://www.futureofprivacy.org/wp-content/uploads/2010/09/Misplaced-Confidences-acquisti-FPF.pdf.
[117] Joseph Turow, Chris Hoofnagle, Deirdre Mulligan, Nathaniel Good & Jens Grossklags, The Federal Trade Commission and Consumer Privacy in the Coming Decade, 3(3) I/S: J. L. & POL'Y FOR INFO. SOC'Y 723 (2007).
[118] Notice that additional factors besides individuals' willingness to participate, such as funding and regional organization, affect the ultimate conversion rate for organ transplants. Hence, Austria, which has an opt-out system, had a deceased organ transplant rate of 20.6 per million people (pmp), whereas the United States, with an opt-in system, had a rate of 26.3 pmp.

launched its News Feed feature in 2006 and had instead waited for users to opt-in,[119] we might not have benefitted from Facebook as we know it today. It is only when data started flowing that users became accustomed to the change. Similarly, had Google solicited consent (or regulatory approval) for "war driving" through cities all over the world to create a comprehensive map of wi-fi networks for its geo-location services, few individuals would agree.[120] Yet in retrospect, after Google provided users an opportunity to opt-out their routers, we suspect that only a negligible number of users have actually done so.[121] The decisions by regulators in this case indicate some appreciation for the value of Google's data use, even if this rationale was not clearly expressed. It is hard to imagine that the continued logging of such data would be permitted by regulators had it not been for an appreciation for Google's very useful geo-location services.

We do not argue that individuals should *never* be asked to expressly consent to the use of their information or offered an option to opt-out. Rather we suggest that the merits of a given data use should be debated as a broader societal issue. Does society believe that direct marketing, behavioral advertising, third party data brokering, or location based services are legitimate (or even commendable) and should be pursued? Or are these excessive intrusions that should be deterred? When making decisions about the need for individuals' consent and how it should be obtained, policymakers should recognize that default rules often prevail and are determinative of the existence of these data uses. Too often debates about whether consent should be solicited or opt-out provided focus solely on the mechanics of expressing consent.[122] But an increasing focus on consent and data minimization, with little appreciation for the value of data use, could jeopardize innovation and beneficial societal advances.

The legitimacy of data use has always been intended to take additional values into account beyond privacy. For example, law enforcement has traditionally been allotted a degree of freedom to override privacy restrictions in appropriate cases.[123] Consequently, the role of consent should be demarcated according to normative choices made by policymakers with respect to prospective data uses. In some cases, consent should not be required; in others, consent should be assumed subject to a right of refusal; in specific cases, consent should be required to legitimize data use.

---

[119] The initial product launch was accompanied by a privacy furor leading Facebook to retract the service, which was rolled out with adjustments. *See* Mark Zuckerberg, An Open Letter from Mark Zuckerberg, THE FACEBOOK BLOG, September 8, 2006, https://blog.facebook.com/blog.php?post=2208562130.

[120] Google's "war driving" featured in a privacy snafu still being investigated by regulators around the globe; yet it concerns the capture by Google of unencrypted payload (content) data – not the practice of mapping wi-fi networks. *See* Kevin O'Brien, European Regulators May Reopen Street View Inquiries, NY TIMES, May 2, 2012, http://www.nytimes.com/2012/05/03/technology/european-regulators-to-reopen-google-street-view-inquiries.html?_r=2.

[121] Kevin O'Brien, Google Allows Wi-Fi Owners to Opt Out of Database, NY TIMES, November 15, 2011, http://www.nytimes.com/2011/11/16/technology/google-allows-wi-fi-owners-to-opt-out-of-database.html.

[122] For extensive discussion *see* Tene & Polonetsky, *supra* note 50.

[123] Law enforcement provisions are also increasingly being limited due to concerns of potential abuse.

The Legal Framework: Solutions

This part argues that while relaxing the principles of data minimization and consent, the current privacy framework should stress access and transparency. It explores how individuals can be empowered with enhanced transparency and access rights, thereby rebalancing the framework and creating additional opportunity for efficient value creation and innovation. It argues that if individuals were provided access to their information in machine-readable usable format, the personal information ecosystem would expand; layers upon layers of user-side applications would emerge to harvest information to benefit not only organizations but also individuals. It further suggests that subject to protection of trade secrets, organizations be required to reveal the criteria used in their decision-making processes with respect to personal data analysis. This will have the effect of discouraging unethical, if not illegal, classifications and provide individuals with the due process opportunity to challenge decisions made about them by algorithm-driven machines.

Access, Portability, and Sharing the Wealth

The right to access and rectify one's individual information, while one of the fundamental principles of information privacy, remains woefully underutilized.[124] Few individuals are aware of their access rights and even fewer exercise them in practice.[125] And why should they? Access rights are neither convenient nor particularly useful. Organizations typically provide access to data only in "hardcopy"; after weeks or months of delays arising from correspondence, requests for authentication and payment of fees; fail to provide details about sources, uses and recipients of information; and seek to rely on a panoply of legal exemptions to mask portions of the data that they do disclose. The increasing complexity of the data ecosystem renders it difficult for individuals to even determine to whom an access request should be sent; and processors or sub-processors of data are often based in foreign jurisdictions without a consumer-facing interface to handle individual requests. Indeed, one user's quest to obtain his personal information from Facebook was so novel that it commanded headlines in newspapers all over the world, including the *New York Times*.[126]

As a *quid pro quo* for looser data collection and minimization restrictions, organizations should be prepared to share the wealth created by individuals' data with those individuals. This means providing individuals with access to their data in machine-readable (we will call it "usable") format; and allowing them to take advantage of applications to analyze their own data and draw useful conclusions (*e.g.*, consume less proteins; go on skiing vacation; invest in bonds).

---

[124] *See, e.g.*, Singer, *supra* note 26.

[125] A Eurobarometer survey of 2008 found that across the EU just over a half of the citizens were aware of the right; far fewer had ever exercised it. Eurobarometer, Data Protection in the European Union Citizens' perceptions Analytical Report, February 2008, p. 30 http://ec.europa.eu/public_opinion/flash/fl_225_en.pdf.

[126] Kevin O'Brien, Austrian Law Student Faces Down Facebook, NY TIMES, February 5, 2012, http://www.nytimes.com/2012/02/06/technology/06iht-rawdata06.html?pagewanted=all.

This "featurization" of big data will unleash innovation and create a market for personal data applications.[127] The technological groundwork has already been completed with mash-ups and real-time APIs making it easier for organizations to combine information from different sources and services into a single user experience. Much like open source software or creative commons licenses, free access to personal data is grounded in both efficiency and fairness rationales. Regardless of whether or not you accept a property approach to personal information,[128] fairness dictates that individuals enjoy beneficial use of their data.

Consider for example the roll out of the smart grid. Electric utilities reap most of the benefits associated with upgrading the electric grid to provide bi-directional communications. This explains why the smart grid was met by pushback from consumers and regulators who are concerned with its implications for privacy, data security, start-up costs and dynamic pricing. Had consumers felt the beneficial impact of the smart grid themselves, they would likely have reacted differently. That is precisely the idea behind the Obama Administration's "Green Button" initiative: that consumers should have access to their own energy usage information in a downloadable, standard, easy-to-use electronic format.[129] In a speech on September 15, 2011, Aneesh Chopra, the U.S. Chief Technology Officer, challenged the industry to "publish information online in an open format (machine readable) without restrictions that would impede re-use".[130] In January 2012, three major California utilities announced their implementation of the Green Button;[131] a dozen more utilities followed suit in the first quarter of 2012.[132]

The Administration predicted that making user data available to the public would lead entrepreneurs to develop technologies like energy management systems and smartphone applications that can interpret and use such information. Homeowners, in turn, would seek out applications that enable them to gain greater control over their energy use. Chopra emphasized the importance of providing the data in a *standard format* according to industry-accepted guidelines. A standard, usable format fosters innovation by allowing software developers to create a single version of their product that will work for all utility customers across the country. One developer told the *New York Times* that: "his company had created a set of software

---

[127] Such a market is already picking up. *See, e.g.*, Francesca Robin, The emerging market that could kill the iPhone, Fortune, August 1, 2012, http://tech.fortune.cnn.com/2012/08/01/iphone.

[128] *See* discussion and criticism of the property approach in Julie Cohen, Examined Lives: Informational Privacy and the Subject as Object, 52 STAN. L. REV. 1373 (2000).

[129] National Institute of Standards and Technology, Green Button Initiative Artifacts Page, http://collaborate.nist.gov/twiki-sggrid/bin/view/SmartGrid/GreenButtonInitiative.

[130] Aneesh Chopra, Remarks to GridWeek, September 15, 2011, http://www.whitehouse.gov/sites/default/files/microsites/ostp/smartgrid09-15-11.pdf.

[131] The utilities are Pacific Gas & Electric, Southern California Edison and San Diego Gas & Electric. See Jim Witkin, Pushing the Green Button for Energy Savings, NY TIMES, January 20, 2012, http://green.blogs.nytimes.com/2012/01/20/a-phone-app-for-turning-down-the-thermostat.

[132] News Release, SGCC Members Lead Industry in Green Button Initiative: Fifteen members of consumer-focused smart grid nonprofit sign on to Green Button, April 25, 2012, http://smartgridcc.org/wp-content/uploads/2012/04/Green-Button-Press-Release.pdf.

development tools that had already attracted 150 app developers. His company also plans to set up an online marketplace, similar to Apple's iPhone App Store or Google's Android Market, where homeowners could download energy-related applications."[133]

Accessing information about energy consumption for cost savings and novel usage is not solely the domain of utilities. For example, the Nest Learning Thermostat, developed by Nest Labs, is an energy conserving, self-programming, slickly designed home thermostat. It is also wi-fi connected to allow users to adjust their home or office temperature via an iPhone or Android app from anywhere they happen to be.[134] Like the Green Button, the Nest Learning Thermostat lets users tap into their own data trail, which includes their movements about the house and information about their daily routine. Major communications providers such as AT&T, Verizon and Comcast have also launched innovative home services focused on energy management and home security and control.[135]

The concept of the "Green Button" follows a path charted by a similar government initiative in the field of health data. In 2010, the Obama Administration announced the "Blue Button", a web-based feature through which patients can easily download their health information in usable format and share it with health care providers and trusted third parties. To make the information more useful, the initiative challenged developers to create applications that build on the Blue Button by helping consumers use their data to manage their own health. In turn, apps such as the Blue Button Health Assistant, developed by Adobe, sprung up to facilitate linkage of patient information, including immunizations, allergies, medications, family health history, lab test results, and more.[136]

An additional government program based on a similar mind-set is the "Data.gov" initiative. Government has long been the biggest generator, collector and user of data (not necessarily PII), keeping records on every birth, marriage and death; compiling figures on all aspects of the economy; and maintaining statistics on licenses, laws and the weather. Until recently, all of the data was locked tight and hard to find even if publicly accessible.[137] In many countries, a freedom of information request to obtain information about the budgetary process, for example, would yield, at best, a voluminous PDF document locked for editing and difficult to explore. The Obama Administration, led by United States Chief Information Officer Vivek

---

[133] *Ibid*.

[134] David Pogue, A Thermostat That's Clever, Not Clunky, NY TIMES, November 30, 2011, http://www.nytimes.com/2011/12/01/technology/personaltech/nest-learning-thermostat-sets-a-standard-david-pogue.html.

[135] *See, e.g.*, Jordan Crook, AT&T Introduces Digital Life: IP-Based Home Automation And Security System With 24/7 Monitoring Centers, TECHCRUNCH, May 7, 2012, http://techcrunch.com/2012/05/06/att-introduces-digital-life-ip-based-home-automation-and-security-system-with-247-monitoring-centers.

[136] Aneesh Chopra, Todd Park & Peter Levin, 'Blue Button' Provides Access to Downloadable Personal Health Data, WHITE HOUSE BLOG, October 7, 2010, http://www.whitehouse.gov/blog/2010/10/07/blue-button-provides-access-downloadable-personal-health-data.

[137] *See, e.g.*, Amanda Conley, Anupam Datta, Helen Nissenbaum & Divya Sharma, Sustaining Privacy and Open Justice in the Transition to Online Court Records: A Multidisciplinary Inquiry, 71 MD. L. REV. 772 (2012).

Kundra, innovated in this sphere by launching "Data.gov". The stated purpose of the new website was "to increase public access to high value, machine-readable datasets generated by the Executive Branch of the Federal Government."[138] The opening of the government's data coffers unleashed a wave of innovation and new economic value, as individuals and businesses used raw data to improve existing services and offer new solutions.[139]

Increased use by individuals of their own data is possible not only in the public sector. Various existing business models seek to arbitrate between users and organizations in order to tilt the scale back in favor of individuals. Harvard Berkman Center's "ProjectVRM" (stands for "vendor relationship management"), which set an admittedly "immodest ambition of turning business on its head", seeks to "provide customers with tools that provide both independence from vendor lock-in and better ways of engaging with vendors – on terms and by means that work better for both sides."[140] In his 2012 book "*The Intention Economy*", ProjectVRM's leader Doc Searls posits a vision of a world where an individual is in complete control of her digital persona and grants permissions for vendors to access it on her own terms. In this world, applications would work for individuals to signal their needs, which vendors would then compete to fulfill.[141]

Personal.com, for example, is a start up enabling individuals to own, control access to and benefit from their personal information.[142] It does so by providing individuals with an online "data vault" divided into compartments called "gems," where they can store and share information about their shopping habits, travel, log-in credentials on various sites, location information, and more.[143] There are currently more than 100 gems with more than 3,000 fields of data. The food preferences gem, for example, includes allergies, religious and dietary restrictions, and whether a user likes spicy food. Users can share gems with family, friends, employees or colleagues, and more important, monetize their own data by selling access to gems to commercial entities. (Personal.com would collect a 10 percent fee on such sales). The

---

[138] The open society: Governments are letting in the light, THE ECONOMIST, February 25, 2010, http://www.economist.com/node/15557477.

[139] *Ibid*.

[140] Harvard University, Berkman Center for Internet and Society, ProjectVRM, http://cyber.law.harvard.edu/research/projectvrm#.

[141] Doc Searls, THE INTENTION ECONOMY: WHEN CUSTOMERS TAKE CHARGE (Harvard Business Review Press, 2012); *also see* Joe Andrieu, Introducing User Driven Services, JOEANDRIEU.COM, April 26, 2009 (series of ten blog posts), http://blog.joeandrieu.com/2009/04/26/introducing-user-driven-services.

[142] Thomas Heath, Web site helps people profit from information collected about them, WASHINGTON POST, June 26, 2011, http://www.washingtonpost.com/business/economy/web-site-helps-people-profit-from-information-collected-about-them/2011/06/24/AGPgkRmH_story.html.

[143] For personal data vaults, *also see* Jerry Kang, Katie Shilton, Deborah Estrin, Jeff Burke & Mark Hansen, Self-Surveillance Privacy, 97 IOWA L. REV. 809 (2012) (also proposing "personal data guardians" to curate the personal data vaults),

company's founders hope that Personal.com will become more than just a data vault, but rather a platform allowing apps to connect to structured user information.[144]

Another example is Intuit's use of data gleaned from its Quickbooks and Turbotax products, which are used by millions of small businesses and individuals for accounting and tax filings. One new feature added to Quickbooks in 2012 is Easy Saver, which looks for items small business owners purchased frequently and then finds a better price for such items using negotiated high-volume discounts. Users will not see an offer for an item unless they have already bought it and are likely (based on previous purchasing behavior) to need it again soon. The Trends feature in Quickbooks tells business owners how their key indicators such as sales, operating margin and payroll costs compare with similar small businesses in their area or in the U.S. overall.[145]

If users fail to exercise their access and rectification rights, why should we expect them to actively engage with their data? The answer is that they are already doing so through a plethora of Apple, Android and Facebook apps.[146] The entire "app economy" is premised on individuals accessing their own data for novel uses, ranging from GPS programs and restaurant recommendations to self-tailored financial and health services.[147] Apps have become an integral aspect of how users experience social networks and the mobile Internet. They enable individuals to make innovative use of their list of friends on Facebook, address books, wi-fi router locations, and many other sources of data. A recent study found that the app economy created 466,000 jobs in the U.S. since 2007.[148] According to Facebook's S-1 filing ahead of its IPO, Zynga, an app developer, is responsible for 12% of Facebook's revenue estimated at more than $4 billion.[149]

We suggest the development of apps for the big data silos of the many companies who have focused on the collection and analysis of personal data for their own use.[150] What the government seeks to achieve with its Green Button and Blue Button initiatives can and should be replicated in the private sector.

The call for additional access and transparency echoes one of the fundamental rationales for information privacy law – the prevention of secret databases. From its inception, information privacy law has been modeled to alleviate the concerns, which arose in the Watergate period in

---

[144] See further discussion in World Economic Forum, Rethinking Personal Data: Strengthening Trust, May 2012, at p. 26-27, https://www.bcgperspectives.com/Images/Rethinking_Personal_Data_1005_light_tcm80-105516.pdf.

[145] Upbin, *supra* note 85.

[146] Michael Liedtke, Study: App economy is a booming jobs engine, USA TODAY, February 7, 2012, http://www.usatoday.com/tech/news/story/2012-02-07/apps-economy-creates-jobs/52997386/1.

[147] *See, e.g.*, Steven Overly, Mobile health apps prompt questions about privacy, WASHINGTON POST, April 29, 2012, http://www.washingtonpost.com/business/capitalbusiness/mobile-health-apps-prompt-questions-about-privacy/2012/04/27/gIQAk17FqT_story.html.

[148] Michael Mandel, Where the Jobs Are: The App Economy, TECHNET, February 7, 2012, http://www.technet.org/wp-content/uploads/2012/02/TechNet-App-Economy-Jobs-Study.pdf.

[149] Anthony Haw, Zynga Makes Up 12 Percent of Facebook's Revenue, TECHCRUNCH, February 1st, 2012, http://techcrunch.com/2012/02/01/zynga-makes-up-12-percent-of-facebooks-revenue.

[150] Singer, *supra* note 26.

the United States and the Communist era in Eastern Europe of secret databases used to curtail individual freedoms.[151] Yet the frameworks that emerged in response to such concerns, providing access rights in the United States and requiring database registration in the European Union, failed to engage individuals who remained largely oblivious to their rights.[152]

The big data model has now reinvigorated the specter of massive data silos accumulating and using information for obscure purposes. Individuals and regulators do not condemn big data as such; rather they oppose "secret big data", which raises a Kafkaesque vision of an inhumane bureaucracy.[153] This means that we must retrofit transparency obligations and access rights to make them more useful in practice. Any activity performed in the dark raises suspicion of being untoward; what is done in broad daylight must be wholesome and "clean".

The call for transparency is not new, of course. Rather the emphasis is on access to data in *usable format*, which can work to create value to individuals. Transparency and access alone have not emerged as potent tools because individuals do not care for, and cannot afford to indulge in transparency and access for their own sake. That is why they seldom opt-in or opt-out of any end user license agreement (EULA) or privacy policy, regardless of their merits. The enabler of transparency and access is the ability to *use* the information and *benefit* from it in a tangible way. This will be achieved through "featurization" or "app-ification" of privacy. Useful access to PII will engage individuals; invite scrutiny of organizations' information practices; and thus expose potential misuses of data. It would be value minimizing to leave this opportunity untapped. Organizations should build as many dials and levers as needed for individuals to engage with their data.

The extent of transparency and access we espouse will no doubt raise serious legal and business complexities. First, organizations (particularly non-consumer facing) may argue that in many circumstances providing individual access to massive databases distributed across numerous servers and containing zetabytes of de-identified data is simply not practical. Second, to avoid the creation of a bigger privacy problem than it seeks to solve, direct online accessibility to data requires strong authentication as well as secure channeling, imposing costs and inconveniences on both organizations and individuals. Third, as the ecosystem for personal information expands, building layers upon layers of user-side applications over the existing centralized structure, so do the risks of data leakage and unauthorized use. Finally, access to machine-readable data in usable format appears to promote data portability, a contentious concept which raises further questions regarding intellectual property and antitrust. Although further work is required to address these concerns, we believe they can be contained.

---

[151] Spiros Simitis, Reviewing Privacy in an Information Society, 135 U. PA. L. REV. 707 (1987); James Whitman, The Two Western Cultures of Privacy: Dignity Versus Liberty, 113 YALE L.J. 1151 (2004); *also see* Michel Foucault, DISCIPLINE AND PUNISH (Vintage Books 1995).

[152] Omer Tene, There is no new thing under the sun, CONCURRING OPINIONS BLOG, July 30, 2012, http://www.concurringopinions.com/archives/2012/07/there-is-no-new-thing-under-the-sun.html.

[153] Dan Solove, THE DIGITAL PERSON: TECHNOLOGY AND PRIVACY IN THE INFORMATION AGE (NYU Press 2004), chapter 3.

First, to be sure, if data were in fact robustly de-identified, it would be counterproductive to require their re-identification simply in order to provide individuals with access.[154] Yet in precisely such circumstances the risk to individuals' privacy would be greatly reduced. Access is most needed where de-identification is weak and the data could therefore provide tangible benefits to individuals. Here too, the flexibility and modularity of the FIPPs proves instrumental: more de-identification, less access rights; less de-identification, more access rights.

Second, privacy and data security clearly require that only a correct individual be granted access to his or her personal data. This means that organizations must authenticate the identity of an individual making a request; and that data are delivered on a secure channel. Implementation may require use of digital signatures and similar identity infrastructures already in existence today; as well as encrypted communication delivery channels.

Third, the enhancement of big data with interfaces for user interaction increases the number of access points and correspondingly the risk of security breach and data leakage.[155] Yet, this is a price worth paying where the goal is data empowerment of individuals. We find unconvincing the argument that individuals should not be allowed to access their information in case it might leak. It is tantamount to a bank barring customers' access to their accounts to avoid their losing their money.

Finally, although with similarity to the data portability argument, our proposal comes short of advocating portability.[156] We recognize that portability is not, strictly speaking, a concept of privacy law but rather one derived from antitrust. It regards personal information as an asset of individuals, which remains under their control unless traded for a fair price. Although the proposed European Data Protection Regulation seeks to weave portability into the fabric of privacy law,[157] we believe this may go too far. The property metaphor fails to capture the psychological and sociological nuance of the right to privacy. As Julie Cohen wrote a decade ago: "Recognizing property rights in personally-identified data risks enabling more, not less, trade and producing less, not more, privacy".[158] Moreover, a right to portability could eviscerate the competitive advantage gained by companies that have invested significant skill and resources to collect, organize and share data in commercially valuable ways, thereby stifling innovation. Companies vying for control of information markets could use it strategically to corner their competitors. Personal information should be regarded as neither an asset of individuals, impinging on business trade secrets and intellectual property rights, nor of business, excluding individuals from benefit. Rather it should be seen as a valuable joint resource; a basis for value creation and innovation.

---

[154] Schwartz & Solove, *supra* note 104.
[155] *See*, in the context of the Facebook ecosystem, Tene, *supra* note 65.
[156] A new right to data portability has been introduced by the European Data Protection Regulation, art. 18.
[157] Proposed European Data Protection Regulation, Art. 18.
[158] Cohen, *supra* note 128, at p. 1391.

Privacy suffers not only when individuals are *unaware* of data practices but also when they are *uninterested* or *disengaged*. In such an environment, regardless of the regulatory mechanisms in place, there will be insufficient checks on data collection and use. Where individuals can access data in a manner that is engaging, useful or valuable, they will give rise to natural checks on inappropriate behavior, thus serving as a useful compliance mechanism for privacy law.

Enhanced Transparency: Shining the Light

Policymakers have always struggled to draw the line for ethical data use.[159] The discussion often revolved around the definition of "sensitive data". Yet any attempt to exhaustively define categories of sensitivity typically fails, given the highly contextual nature of personal information. For example, the first data protection case taken by the European Court of Justice, the matter of *Bodil Lindqvist*,[160] dealt with the use of "sensitive" information so benign so as to appear trivial – the fact that the defendant's fellow churchgoer had a broken leg. A broken leg is clearly a medical condition, which is a category of sensitive data under any legal framework;[161] yet few would consider it sensitive in fact.

In order to delimit the zone of ethical data analysis we propose that organizations reveal not only the existence of their databases but also the *criteria* used in their decision-making processes, subject to protection of trade secrets and other intellectual property.[162] Today, such disclosures are made only when a user is presented with a consumer privacy policy, and even then the logic behind some of the automated processes remains opaque. Louis Brandeis, who together with Samuel Warren "invented" the legal right to privacy in 1890,[163] has also written that "[s]unlight is said to be the best of disinfectants".[164] We trust if the existence and uses of databases were visible to the public, organizations would be more likely to avoid unethical or socially unacceptable uses of data. If organizations were required to disclose their line of reasoning in data processing operations impacting individuals' lives, they would avoid unethical uses of data concerning, for example; children; legally suspect categories, such as gender, age or

---

[159] See boyd & Crawford, *supra* note 19, at p. 11, stating: "Very little is understood about the ethical implications underpinning the Big Data phenomenon".

[160] Bodil Lindqvist (Case C-101/01) [2003] ECR I-12971.

[161] *See, e.g.*, European Data Protection Directive, art. 8(1); FTC Final Report, sec. IV.C.2.e.ii.

[162] *See* Article 12 of the European Data Protection Directive, which requires organizations to provide an individual with "knowledge of the logic involved in any automatic processing of data concerning him at least in the case of the automated decisions". Recital 41 of the European Data Protection Directive acknowledges the need to protect organizational assets: "every data subject must also have the right to know the logic involved in the automatic processing of data concerning him, at least in the case of the automated decisions (…) this right must not adversely affect trade secrets or intellectual property and in particular the copyright protecting the software (…) these considerations must not, however, result in the data subject being refused all information".

[163] Samuel Warren & Louis Brandeis, The Right to Privacy, 4 HARV. L. REV. 193 (1890).

[164] Louis Brandeis, What Publicity Can Do, HARPER'S WEEKLY, December 20, 1913, http://c0403731.cdn.cloudfiles.rackspacecloud.com/collection/papers/1910/1913_12_20_What_Publicity_Ca.pdf.

race; or sensitive data (in the parochial sense), such as sexual preferences or certain (but not all) medical conditions.

More broadly, the requirement that organizations reveal their decisional criteria is based on the FIPPs' transparency and accuracy principles. In a big data world, what calls for scrutiny is often not the accuracy of the *raw data* but rather the accuracy of the *inferences* drawn from the data. Inaccurate, manipulative or discriminatory conclusions may be drawn from perfectly innocuous, accurate data. Much like in quantum physics, the observer in big data analysis can affect the results of her research by defining the data set, proposing a hypothesis or writing an algorithm. At the end of the day, big data analysis is an interpretative process, in which one's identity and perspective informs one's results. Like any interpretative process, it is subject to error, inaccuracy and bias.[165]

The requirement that organizations disclose their decisional criteria (not necessarily the *algorithms*, but rather the *factors* that figure in them) highlights an important fault line between law and technology. Fairness and due process mandate that individuals know the basis for decisions affecting their lives, particularly those made by machines operating under opaque criteria. In the landmark *Daubert* case, the Supreme Court charged trial judges with the responsibility of acting as gatekeepers to exclude unreliable scientific expert testimony.[166] Following *Daubert*, Justice Scalia remarked in *Melendez-Diaz* that "[f]orensic evidence is not uniquely immune from the risk of manipulation".[167] This was in response to the government's assertion that "there is a difference, for Confrontation Clause purposes, between testimony recounting historical events, which is 'prone to distortion or manipulation,' and the testimony at issue here, which is the 'resul[t] of neutral, scientific testing'".[168] We argue that not only the accused, but also any other citizen be afforded a right to confront decisions made about her. *Daubert* and its progeny mandate that at the end of the day it will be lawyers and judges, not technology, who try individuals.[169]

The rule we propose also focuses regulatory attention on the decision-makers who draw conclusions from personal information rather than other parties in the ecosystem. In doing so, it recognizes that some of the risks of big data affect fairness, equality and other values, which may be no less important than – but are theoretically distinct from – core privacy interests. Over the past few years, the debate over privacy has become conflated with broader social values. For example, the increasing tendency of employers to use social networking services to run background checks on prospective job candidates has led critics to condemn the "privacy

[165] boyd & Crawford, *supra* note 19, at p. 7.
[166] Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993); FED. R. EVID. 702, 28 U.S.C. Paul Giannelli, Daubert and Forensic Science: The Pitfalls of Law Enforcement Control of Scientific Research, 2011 U. ILL. L. REV. 53.
[167] Melendez-Diaz v. Massachusetts, 129 S. Ct. 2527, 2536 (2009).
[168] *Ibid*, id.
[169] *Cf.* Randall Stross, The Algorithm Didn't Like My Essay, NY TIMES, June 9, 2012, http://www.nytimes.com/2012/06/10/business/essay-grading-software-as-teachers-aide-digital-domain.html.

invasive" nature of such platforms. Yet on closer scrutiny, it is not clear that social networking services should be held accountable for illegal or unethical discrimination by employers. If an employer chooses to screen out job candidates based on race, good looks,[170] or proclivity to drink,[171] then that employer – not the neutral platform used to convey such information – should stand to blame. Accordingly, it is prospective employers – or, in other contexts, insurers, banks and government agencies[172] – that need to explain their decisional criteria in reaching personal data driven conclusions.

Finally, attention must be given to the accessibility of big data sets to the research community at large.[173] Traditionally, when scientists published their research, they also made the underlying data available so that other scientists could verify the results. Yet with big data, it is often only the employees of certain organizations that benefit from access, conducting analysis and publishing results without making the underlying data publicly available.[174] Such scientists may argue, first, that the data are a proprietary asset of their business. Indeed, they may claim that disclosing the data could infringe customers' privacy.[175] Who gets access to big data sets; for what purposes; in what contexts; and with what constraints – these are fundamental questions that must be addressed by future research.[176] Without good answers, we may witness a stratification of the scientific world to have – and have-nots – of big data.[177]

Conclusion

Privacy advocates and data regulators increasingly decry the era of big data as they observe the growing ubiquity of data collection and increasingly robust uses of data enabled by powerful processors and unlimited storage. Researchers, businesses and entrepreneurs equally

[170] *See, e.g.*, Attractiveness discrimination: Hiring hotties, THE ECONOMIST, July 21, 2012, http://www.economist.com/node/21559357.
[171] Jeffrey Rosen, The Web Means the End of Forgetting, NY TIMES, 21 July 2010, http://www.nytimes.com/2010/07/25/magazine/25privacy-t2.html.
[172] *See, e.g.*, Omer Tene, What Happens Online Stays Online: Comments on "Do Not Track", STANFORD CIS BLOG, March 26, 2011, http://cyberlaw.stanford.edu/blog/2011/03/what-happens-online-stays-online-comments-do-not-track.
[173] *See* John Markoff, Troves of Personal Data, Forbidden to Researchers, NY TIMES, May 21, 2012, http://www.nytimes.com/2012/05/22/science/big-data-troves-stay-forbidden-to-social-scientists.html.
[174] *See* Lev Manovich, Trending: The Promises and the Challenges of Big Social Data, in DEBATES IN THE DIGITAL HUMANITIES (Matthew Gold, Editor, The University of Minnesota Press, 2012), http://www.manovich.net/DOCS/Manovich_trending_paper.pdf, claiming that "only social media companies have access to really large social data – especially transactional data. An anthropologist working for Facebook or a sociologist working for Google will have access to data that the rest of the scholarly community will not".
[175] *See* Bernardo Huberman, Sociology of Science: Big Data Deserve a Bigger Audience, 482 NATURE 308 (2012), warning that privately held data was threatening the very basis of scientific research, and complaining that "[m]any of the emerging 'big data' come from private sources that are inaccessible to other researchers. The data source may be hidden, compounding problems of verification, as well as concerns about the generality of the results".
[176] boyd & Crawford, *supra* note 17, at p. 12.
[177] *Ibid*, at p. 13, calling this "the Big Data rich and the Big Data poor".

vehemently point to concrete or anticipated innovations that may be dependent on the default collection of large data sets. We call for the development of a model where the benefits of data for organizations and researchers are shared with individuals. If organizations provide individuals with access to their data in usable format, creative powers will be unleashed to provide users with applications and features building on their data for new innovative uses. In addition, transparency with respect to the logic underlying organizations' data processing will deter unethical, sensitive data use and allay concerns about inaccurate inferences. Traditional transparency and individual access mechanisms have proven ineffective to energize individuals to engage with their data. The promise of new benefits and value sharing propositions will incentivize individuals to act without compromising organizations' ability to harness big data.